混合异构数据的清洗、存储、挖掘架构选型和设计策略

@卢亿雷 From AdMaster

johnlya@163.com

提纲

- 混合异构数据特点
- 混合异构数据分类
- 混合异构处理流程
- AdMaster混合异构数据平台架构
- AdMaster数据处理流程
- AdMaster混合异构数据分析
- Q/A

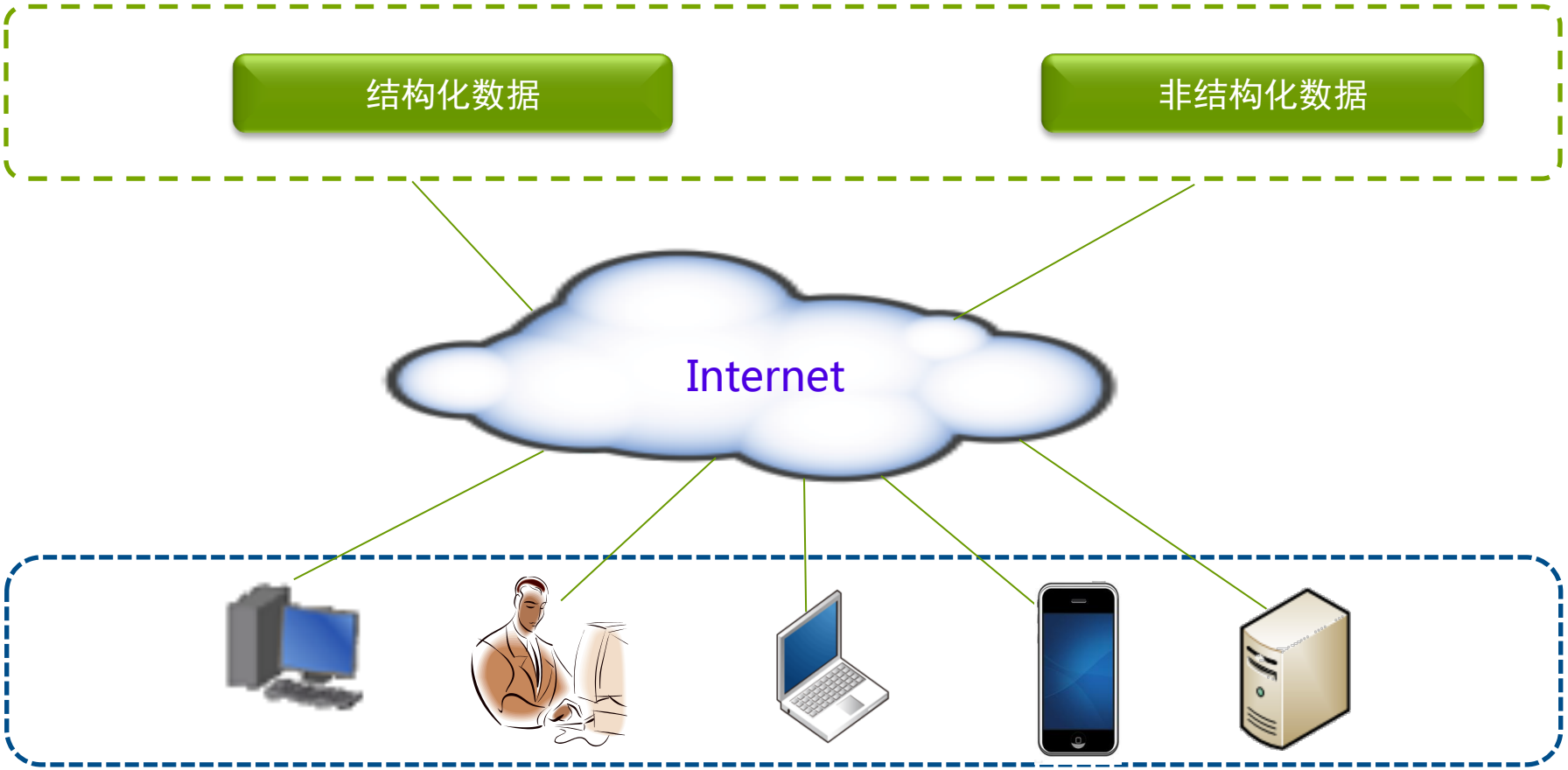
混合异构数据特点

- 不同的数据类型
- 不同的数据量级
- 不同的访问速度
- 不同的用户类型
- 不同的访问平台
- 不同的存储设备
- 。。。

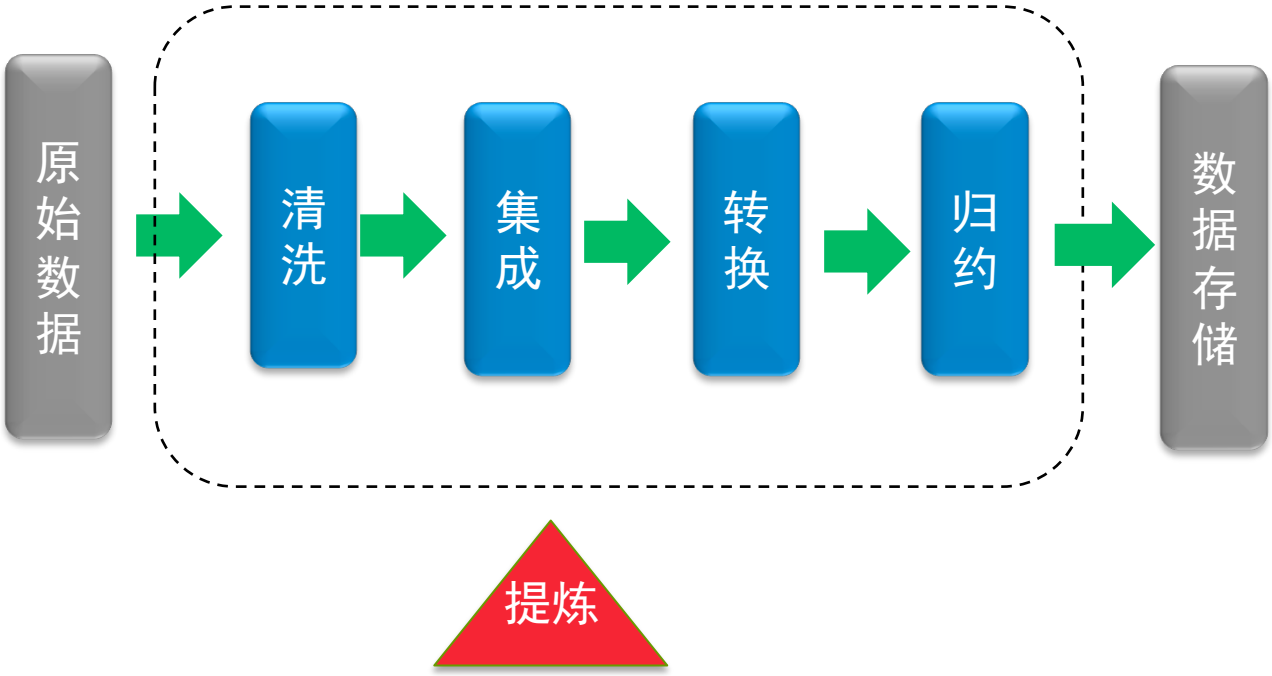
混合异构数据分类

	在线数据	离线数据
数据内容	短周期数据	长周期（存档、归纳、 计算结果）
数据特性	字段固定	字段不固定
数据结构	高度结构化、复杂、适合操作计算	结构简单
使用频率	非常高（热数据）	一般（冷数据）
数据访问量	KB、MB级	GB、TB、PB级
响应时间	纳秒、微秒、毫秒级	秒、分钟、小时、天级

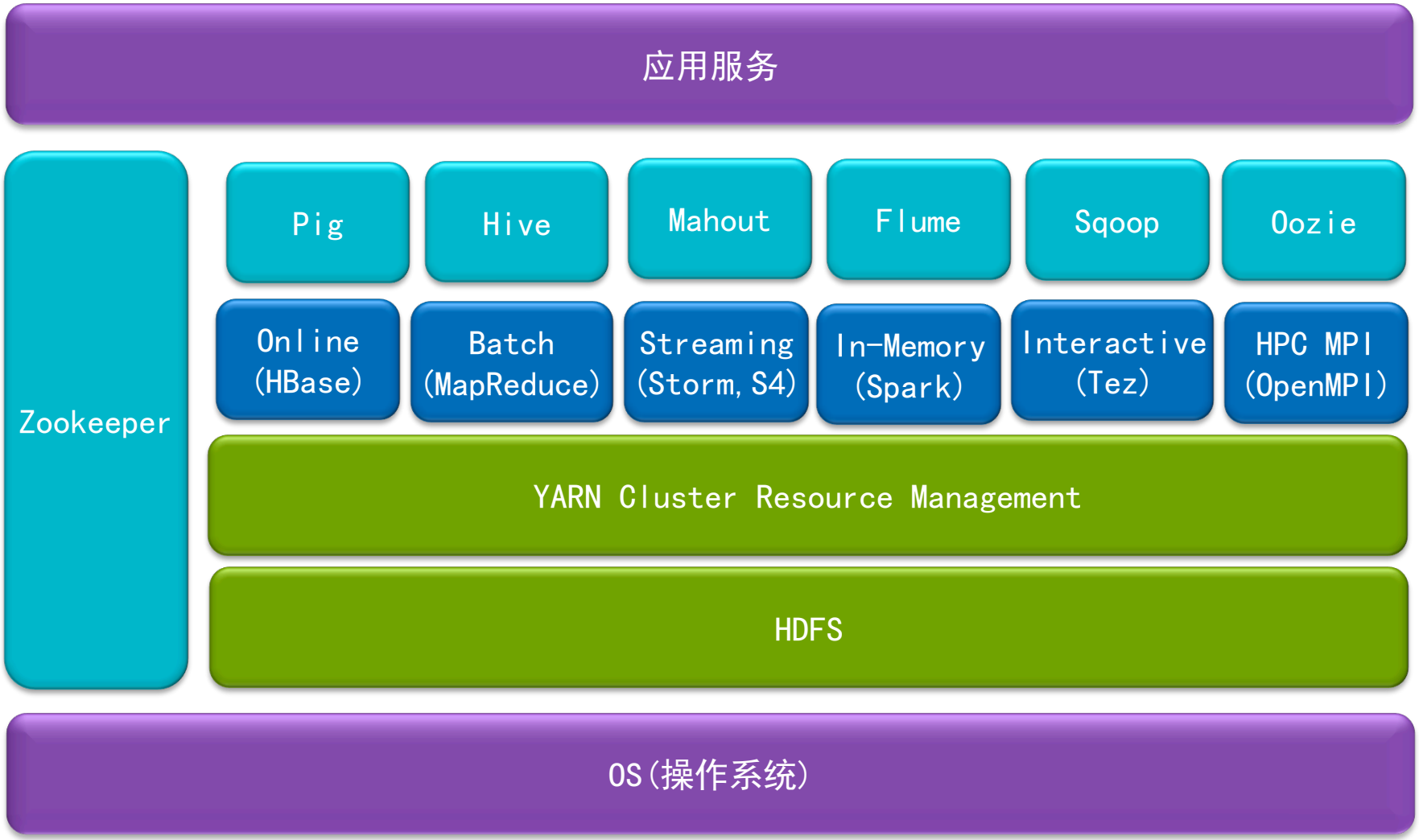
数据采集



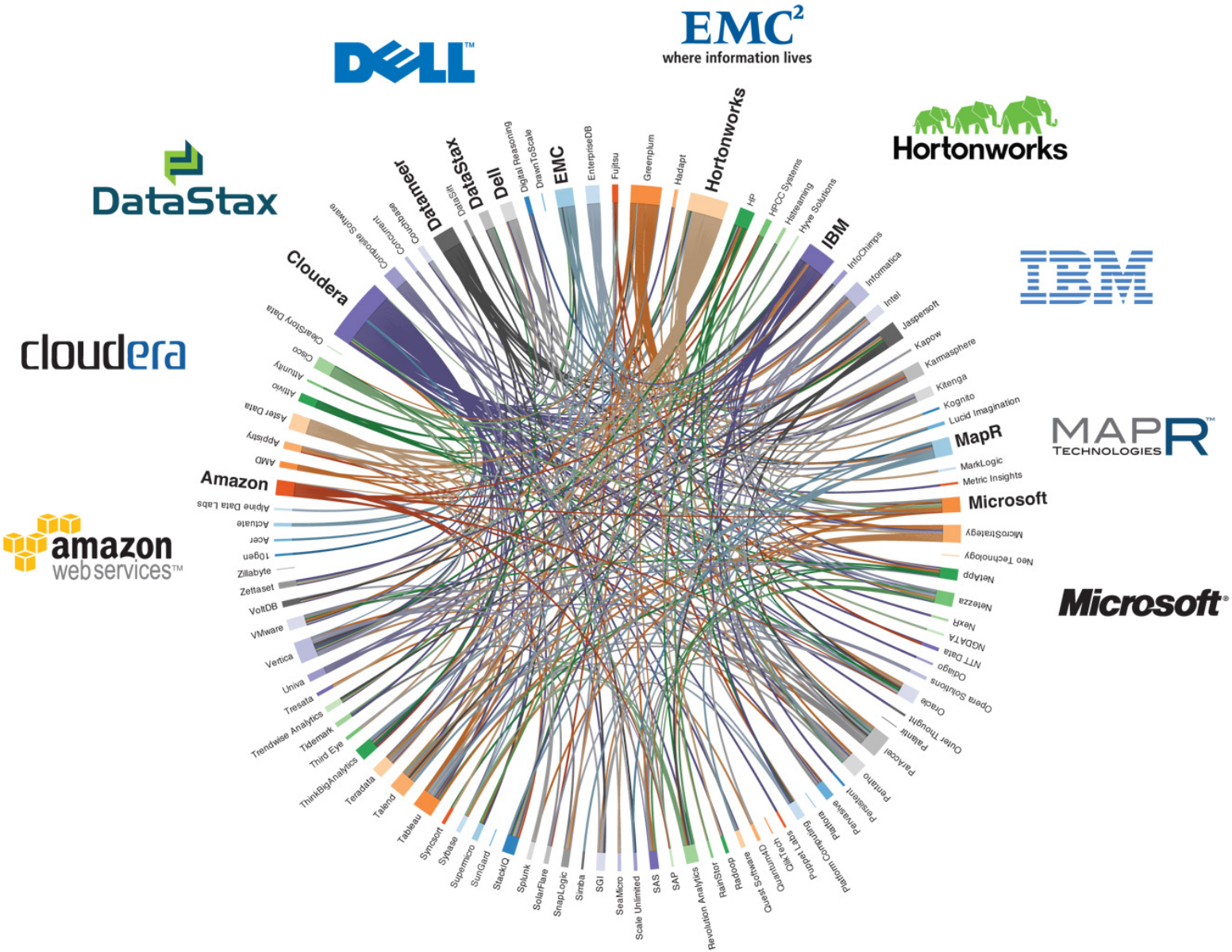
数据预处理



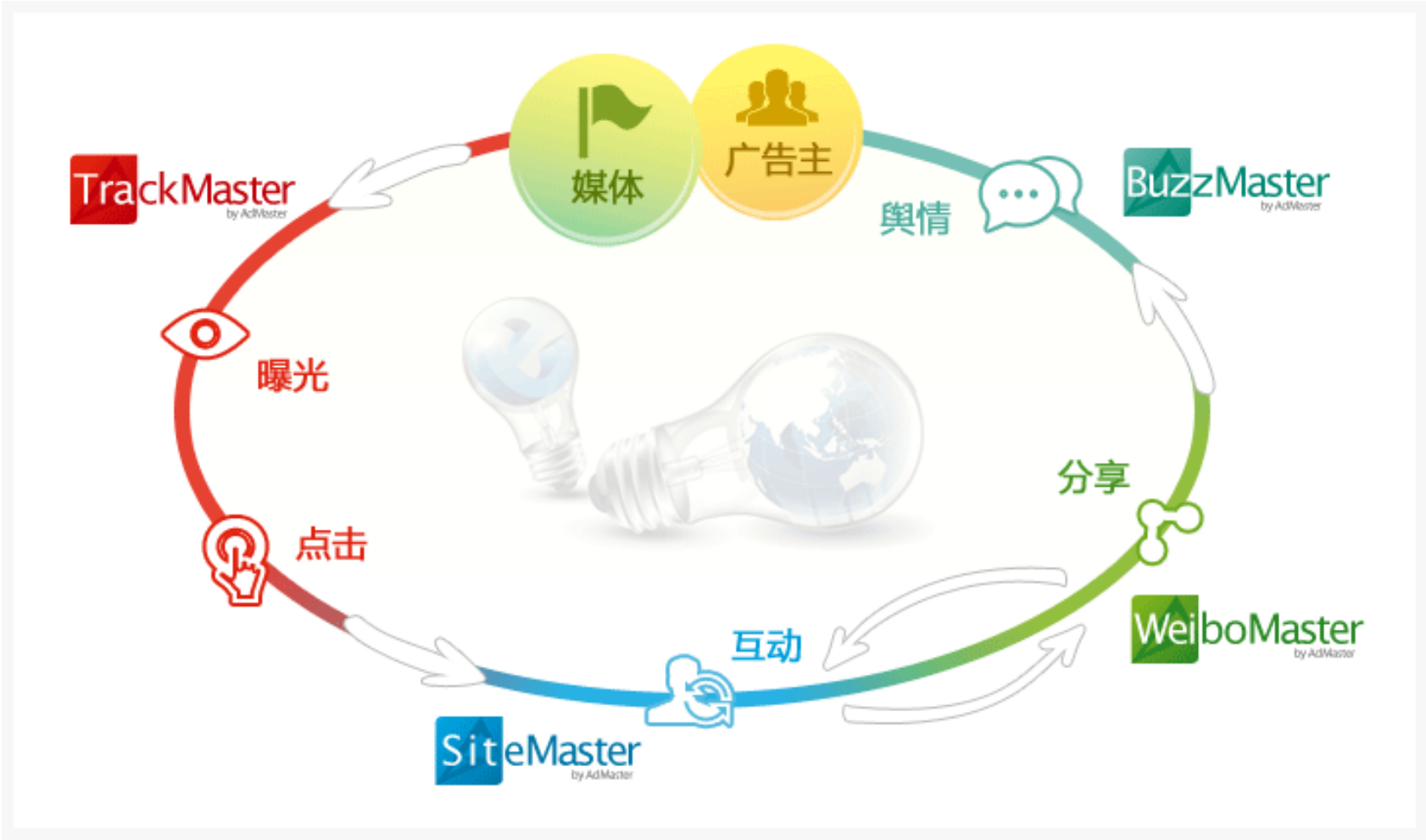
数据分析



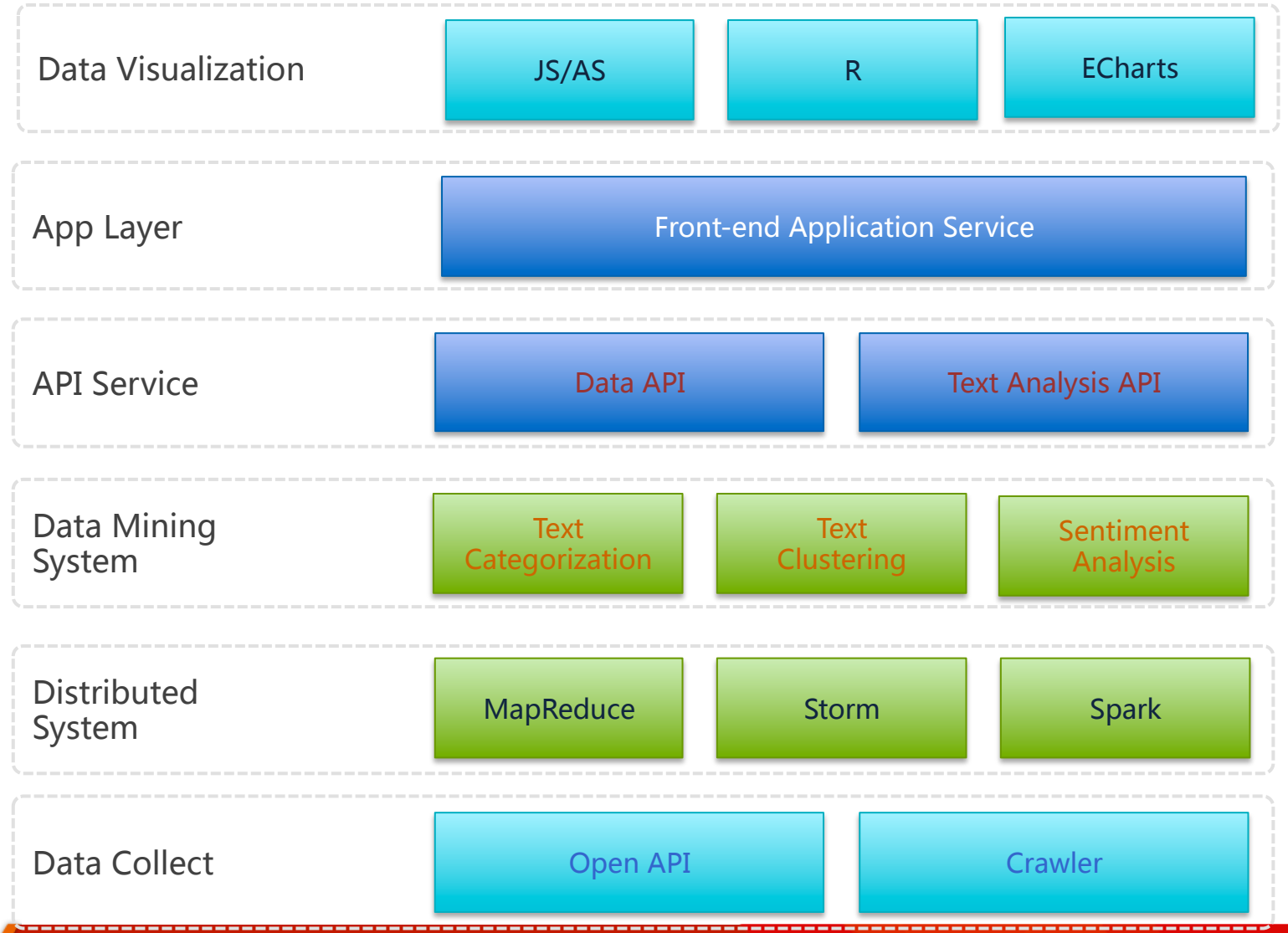
数据展示


DELL
EMC²
where information lives
DataStax
Hortonworks
cloudera
IBM
MAPR
TECHNOLOGIES
amazon
web services™
Microsoft
Hortonworks
DataStax
Dell
EMC
Cloudera
Amazon
IBM
MapR
Microsoft
Datameer
Hadapt
Greenplum
Fujitsu
EnterpriseDB
Digital Reasoning
DrawnToScale
HP
HPCC Systems
Hstreaming
Hyve Solutions
InfoChimps
Informatica
Intel
Jaspersoft
Kapow
Karmasphere
Kitenga
Kognito
Lucid Imagination
MarkLogic
Metric Insights
Microstrategy
Neo Technology
NetApp
Netezza
NeXR
NGDATA
NTT Data
Oracle
Opera Solutions
Outer Thought
Paraccel
Pentaho
Pervasive
Platfora
Platform Computing
Pupper Labs
QlikTech
Quantum10
Quest Software
Radoop
RainStor
Revolution Analytics
SAP
SAS
Scale Unlimited
SeaMicro
SGI
Simba
SnapLogic
SolarFlare
Splunk
StackIQ
SunGard
Supermicro
Sybase
Syncsort
Tableau
Talend
Teradata
ThinkBigAnalytics
Third Eye
Tidemark
Trendwise Analytics
Tresata
Univa
Vertica
VMware
VoltDB
Zettaset
Zillabyte
10gen
Acer
Actuate
Alpine Data Labs
AMD
Appistry
Aster Data
Attivio
Attunity
Cisco
ClearStory Data
Composite Software
Concurrent
Couchbase
DataSift

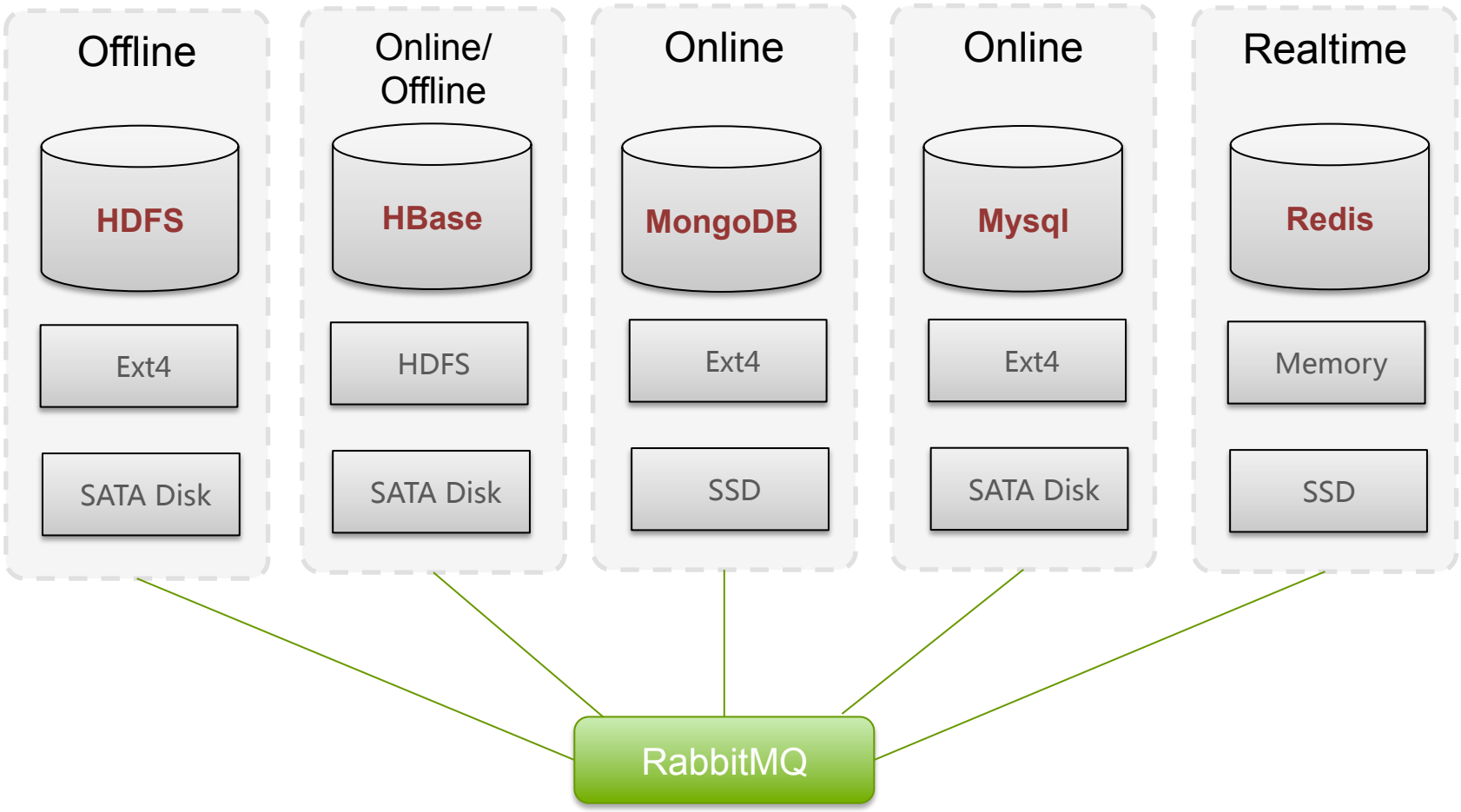
互联网广告监测全流程


媒体
广告主
舆情
BuzzMaster
by AdMaster
分享
WeiboMaster
by AdMaster
互动
SiteMaster
by AdMaster
点击
曝光
TrackMaster
by AdMaster

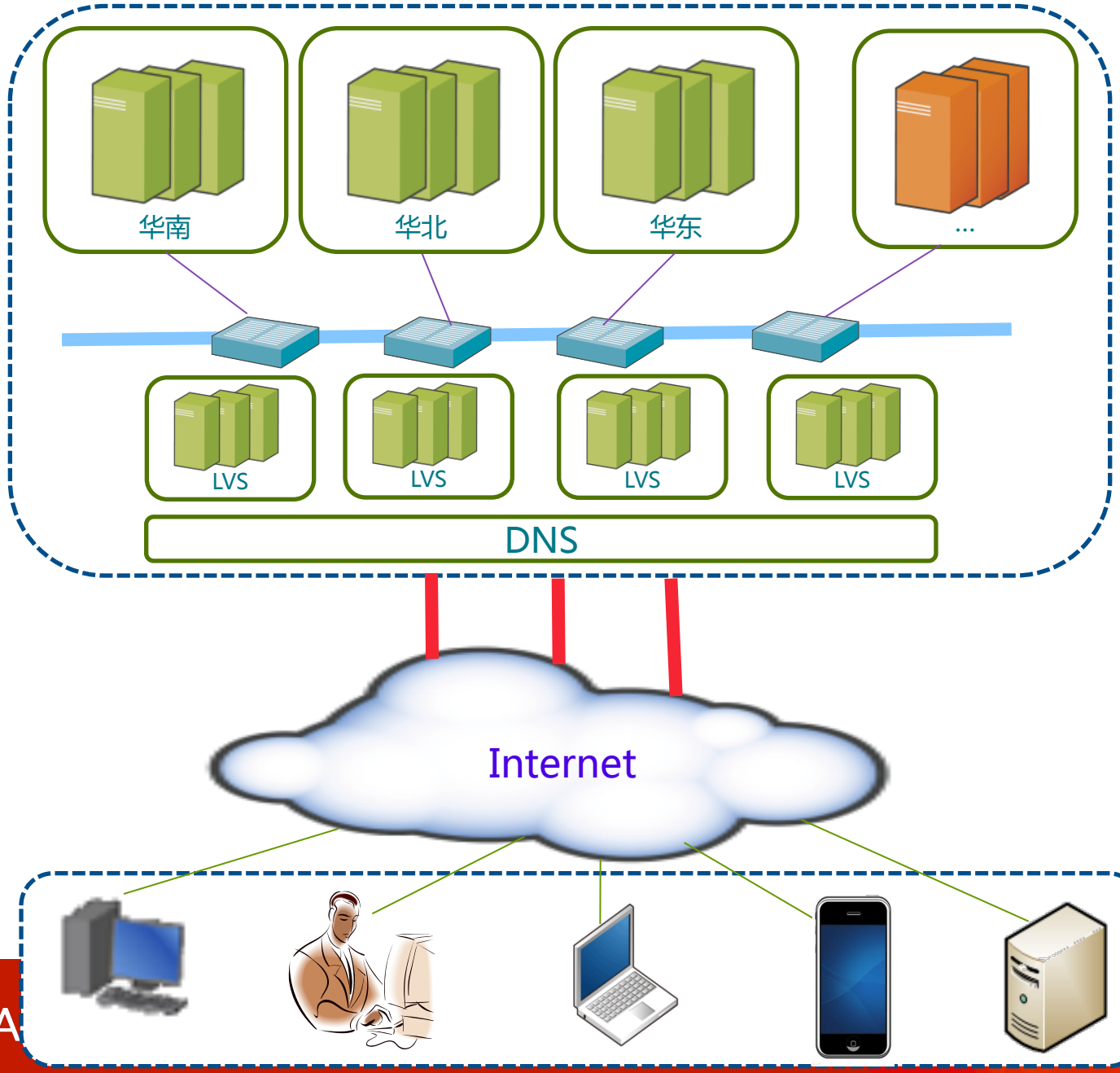
AdMaster混合异构数据平台架构



AdMaster混合异构数据平台架构



AdMaster数据采集



广告数据采集

P&G
Mercedes-Benz
可口可乐
MINI
STARBUCKS
Microsoft

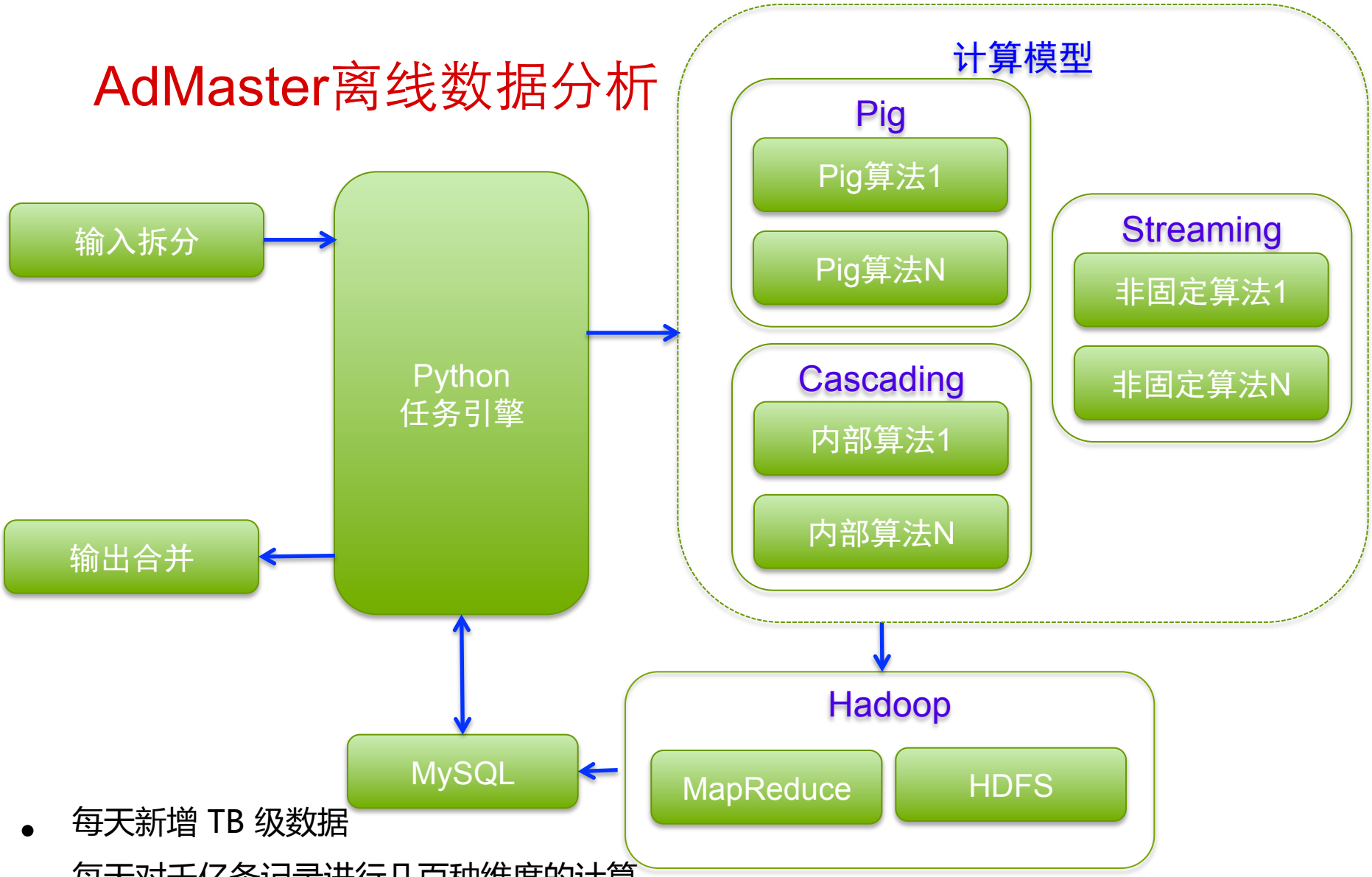
Buzz Resource

网易 NetEase www.163.com
腾讯网 qq.com
搜狐 SOHU.com
sina 新浪
天涯论坛 bbs.tianya.cn
猫扑 mop.com
社区动力 DISCUZ!
Others ...

AdMaster数据采集

- cat /proc/sys/net/ipv4/tcp_mem
- cat /proc/net/sockstat
- cat /proc/sys/net/ipv4/tcp_max_orphans
- net.netfilter.nf_conntrack_max
- net.netfilter.nf_conntrack_tcp_timeout_established

AdMaster离线数据分析

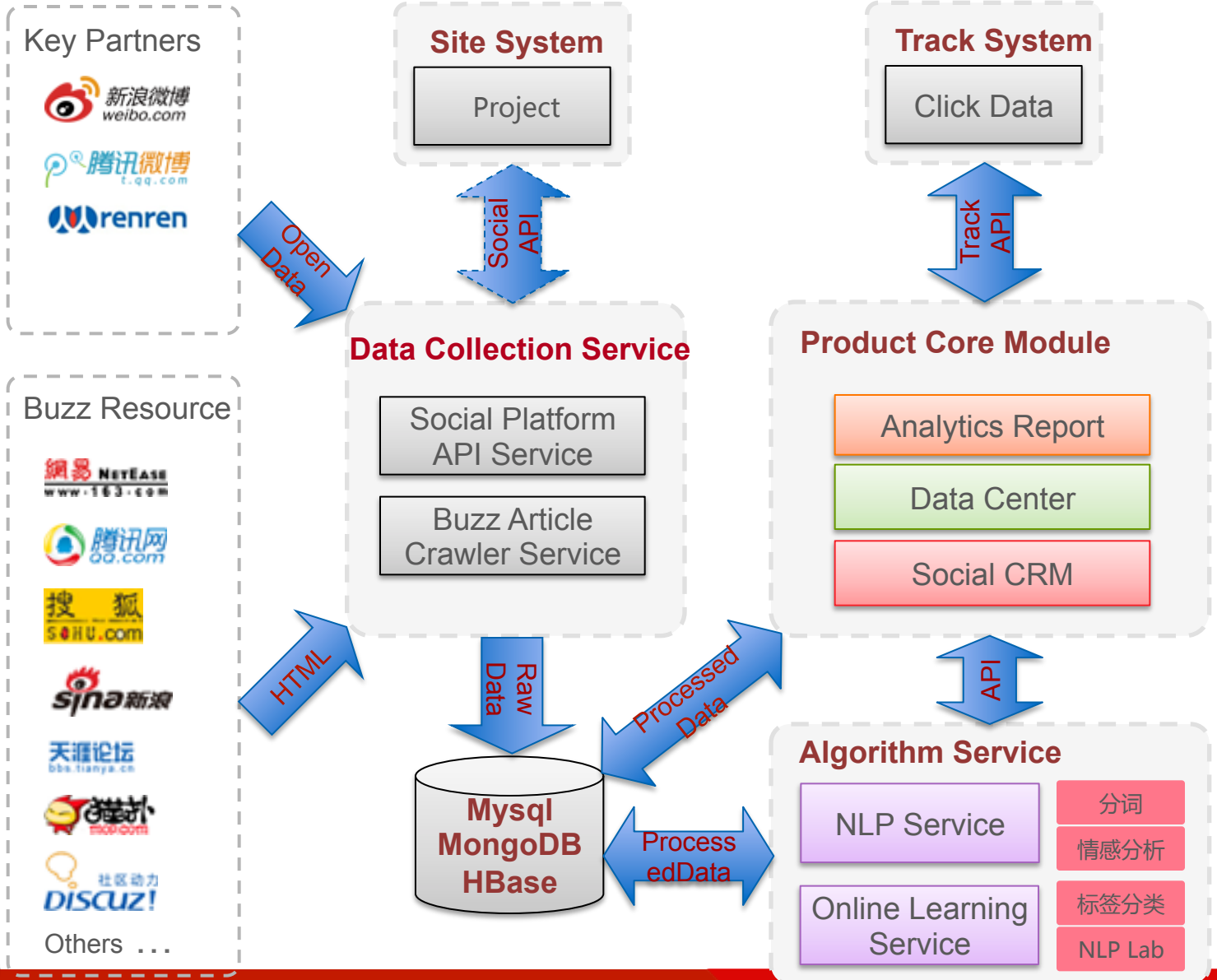


- 每天新增 TB 级数据
- 每天对千亿条记录进行几百种维度的计算

AdMaster离线数据分析

- /sys/kernel/mm/redhat_transparent_hugepage/enabled
- /sys/kernel/mm/redhat_transparent_hugepage/defrag
- dfs.socket.timeout
- dfs.datanode.max.xcievers
- dfs.datanode.socket.write.timeout
- dfs.namenode.handler.count

AdMaster在线数据分析



AdMaster在线数据分析

- Kafka & Tail
- HBase & MongoDB
- Storm & Rsync
- Spark & Hana

AdMaster数据可视化

- 主题 (配色、品牌名、品牌logo)
- 轮播信息 (Screens、Slides、标题)
- 权限 (用户、用户组)

- Social数据源
- Site数据源
- Track数据源

配置信息

数据源

用户请求 → 配置API → 中转服务器API → 数据源API → 前端渲染

AdMaster数据可视化

- 数据呈现方式
- 展示层与数据层松耦合，多种数据源接入
- 极高的可靠性和容错机制

Q & A