实时大数据分析之利器Druid

欧阳辰

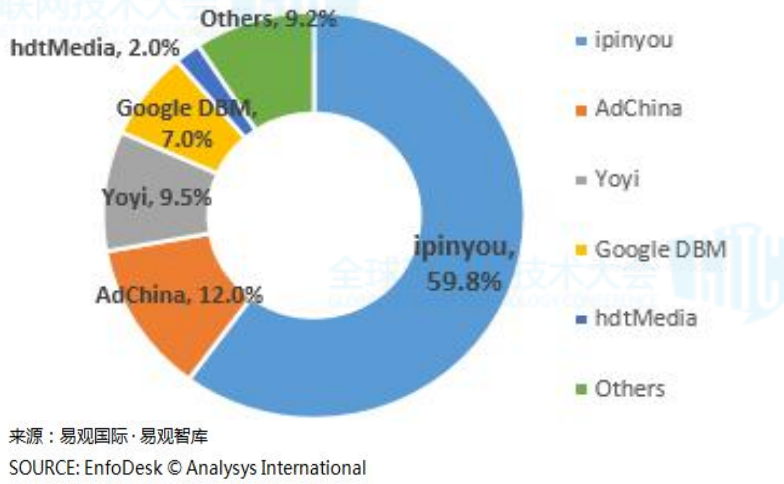
2017/12

议程

- 关于品友
- 大数据分析的繁花似锦
- 历史和发展
- 架构
- 技术优势
- 应用
- 其他分析工具

品友：中国程序化营销的领跑者

品牌程序化市场占有率 59.8%



来源：易观国际·易观智库

SOURCE: EnfoDesk © Analysys International

独立第三方广告技术领先者：创新&执行力

品友大数据计算平台

1.5P

每日处理1P的数据量

260亿

每日处理260亿条日志

8.9亿

8.9亿Cookies人群

4亿

每日覆盖4亿个网页

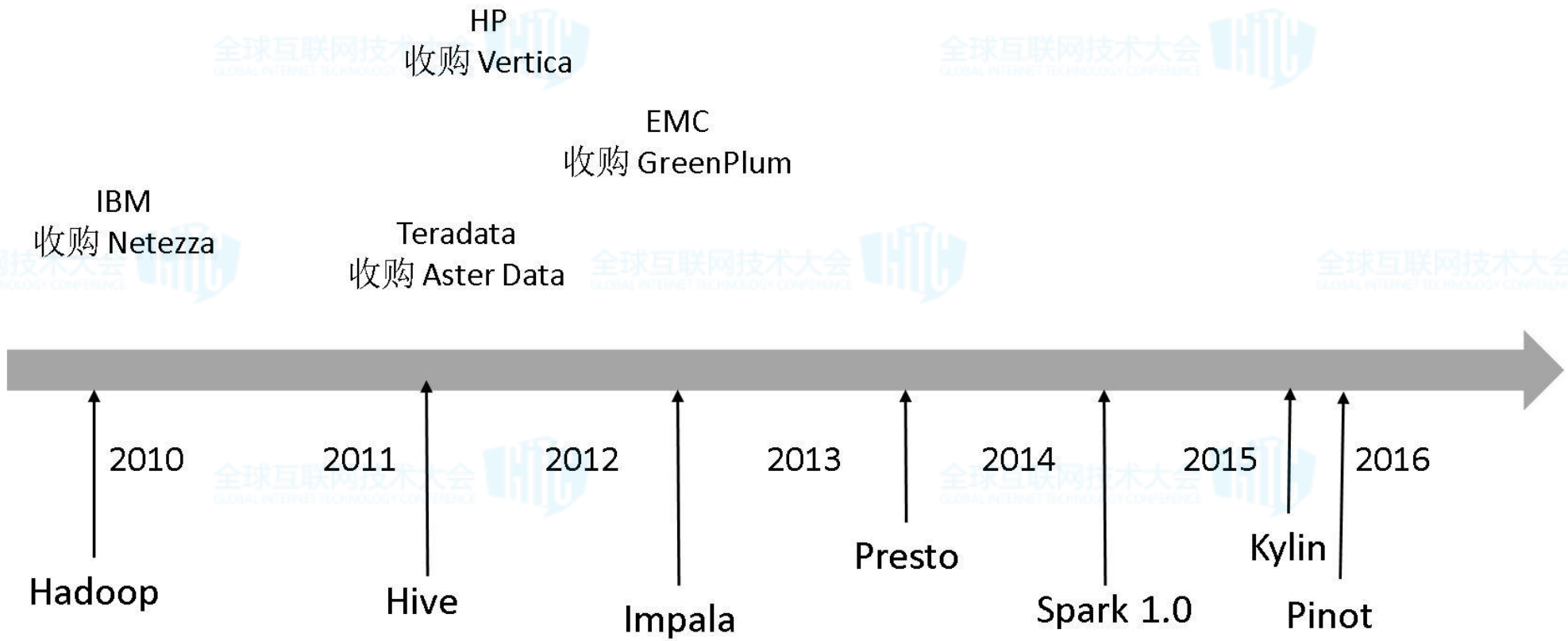
20T

每日20T新增日志

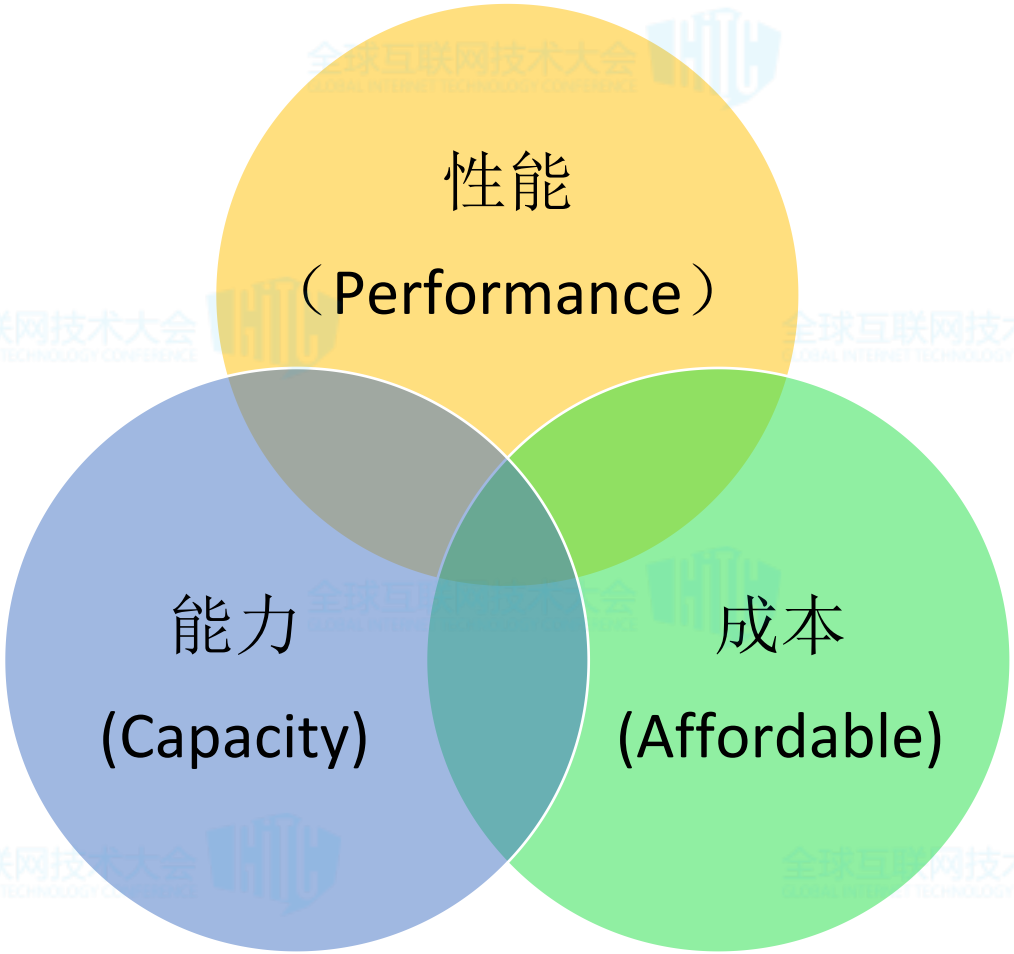
5.3亿

5.3亿日独立用户设备

大数据分析的繁花似景


HP
收购 Vertica
EMC
收购 GreenPlum
IBM
收购 Netezza
Teradata
收购 Aster Data
2010
2011
2012
2013
2014
2015
2016
Hadoop
Hive
Impala
Presto
Spark 1.0
Kylin
Pinot

大数据分析的CAP



例如：

- Druid: A,P
- Vertica: C,P
- Presto: C
- ClickHouse: P, C, A?

DRUID介绍

- 2011 Metamarket 开发，2012年开源
- 初始用于广告分析，程序化分析
- +150贡献者
- 典型应用
  - 300亿事件 /天 (品友互动)
  - 10亿事件/分钟（ Jolata）
  - 用户行为分析（今日头条）
  - 广告实时分析（小米）
  - 性能监控分析（OneAPM）
  - 等等

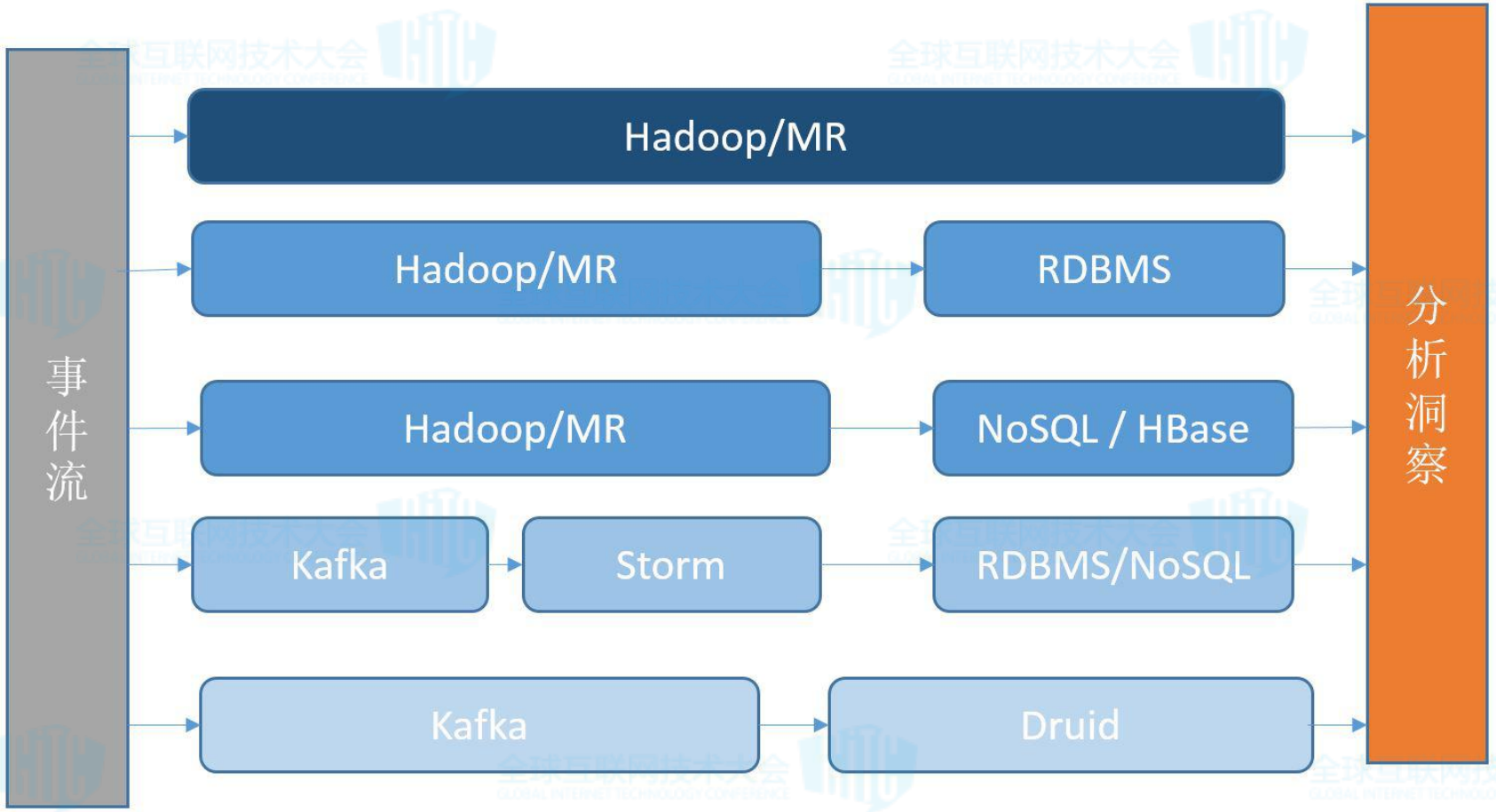
2017, 8, Druid 中国 第五次Meetup

2017, 3, Druid 中国 第四次Meetup

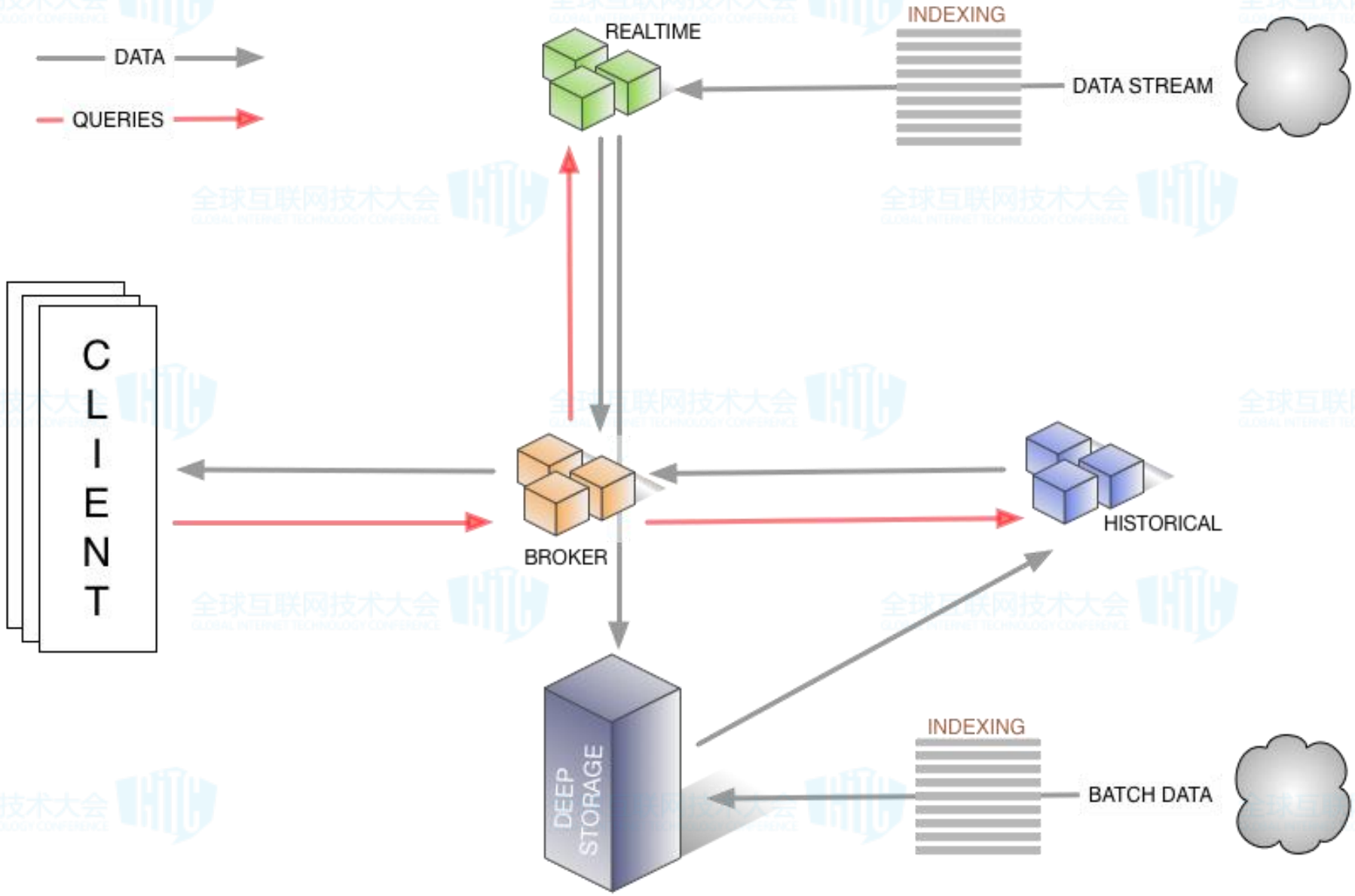
Druid简介

- 高可用性，Segment Shard机制
- 高性能，亚秒级查询响应
- 高吞吐，支持实时数据接入，批量数据接入
- 正确性，lambda架构能够在T+1时间校正实时数据
- 查询有segment级别缓存
- 堆外内存复用，避免GC问题

数据分析的演化阶段


事件流
Hadoop/MR
Hadoop/MR
RDBMS
Hadoop/MR
NoSQL / HBase
Kafka
Storm
RDBMS/NoSQL
Kafka
Druid
分析洞察

DRUID 架构


DATA
QUERIES
REALTIME
INDEXING
DATA STREAM
CLIENT
BROKER
HISTORICAL
DEEP STORAGE
INDEXING
BATCH DATA

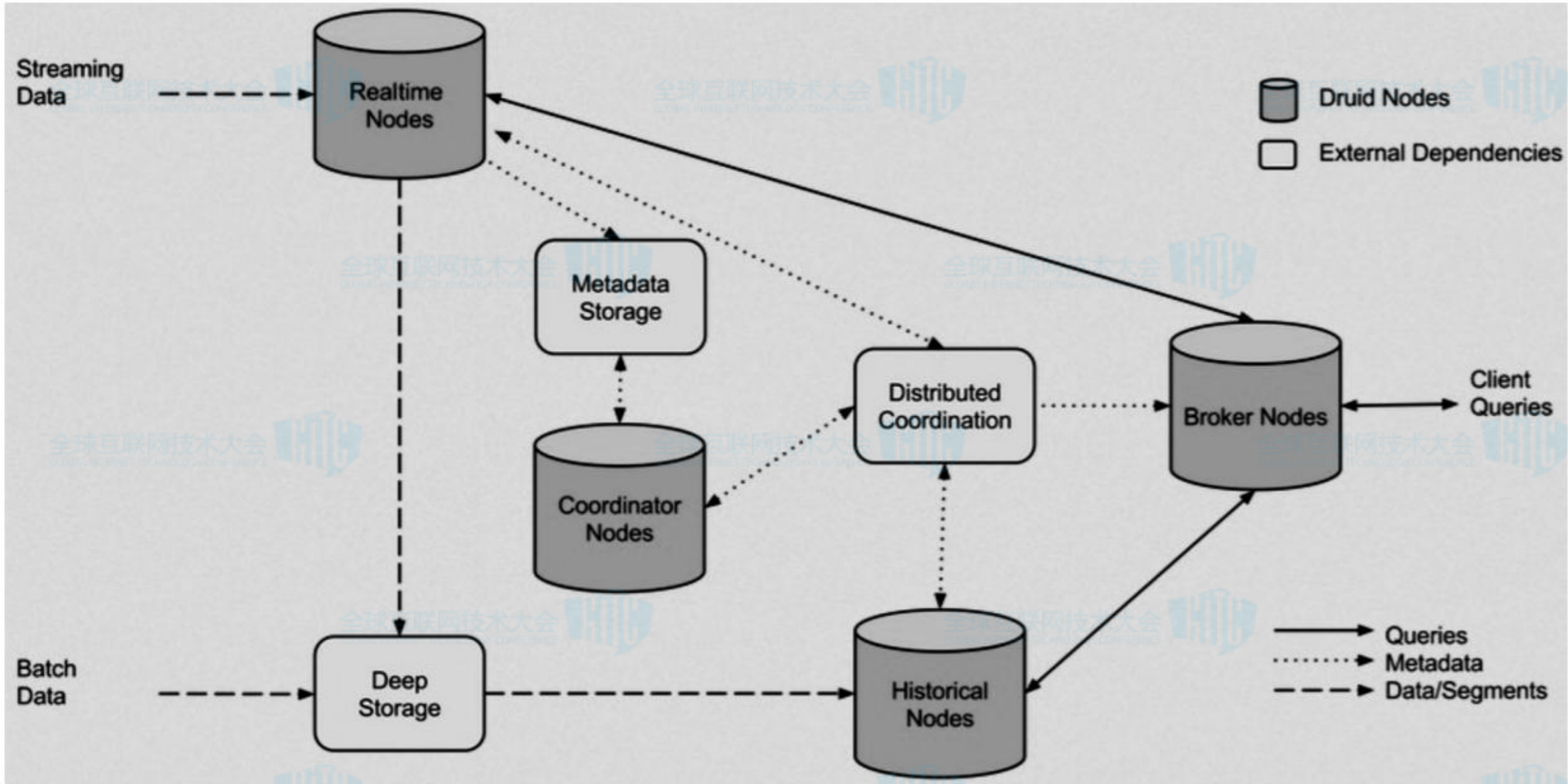


图 3-1　Druid 总体架构图

图 3-10　DataSource 结构

timestamp	publisher	advertiser	gender	country	impressions	clicks	revenue
2011-01-01T01:00:00Z	ultratrimfast.com	google.com	Male	USA	1800	25	15.70
2011-01-01T01:00:00Z	bieberfever.com	google.com	Male	USA	2912	42	29.18
2011-01-01T02:00:00Z	ultratrimfast.com	google.com	Male	UK	1953	17	17.31
2011-01-01T02:00:00Z	bieberfever.com	google.com	Male	UK	3194	170	34.01

图 3-11　DataSource 聚合后的数据情况

Druid 的类 LSM-tree

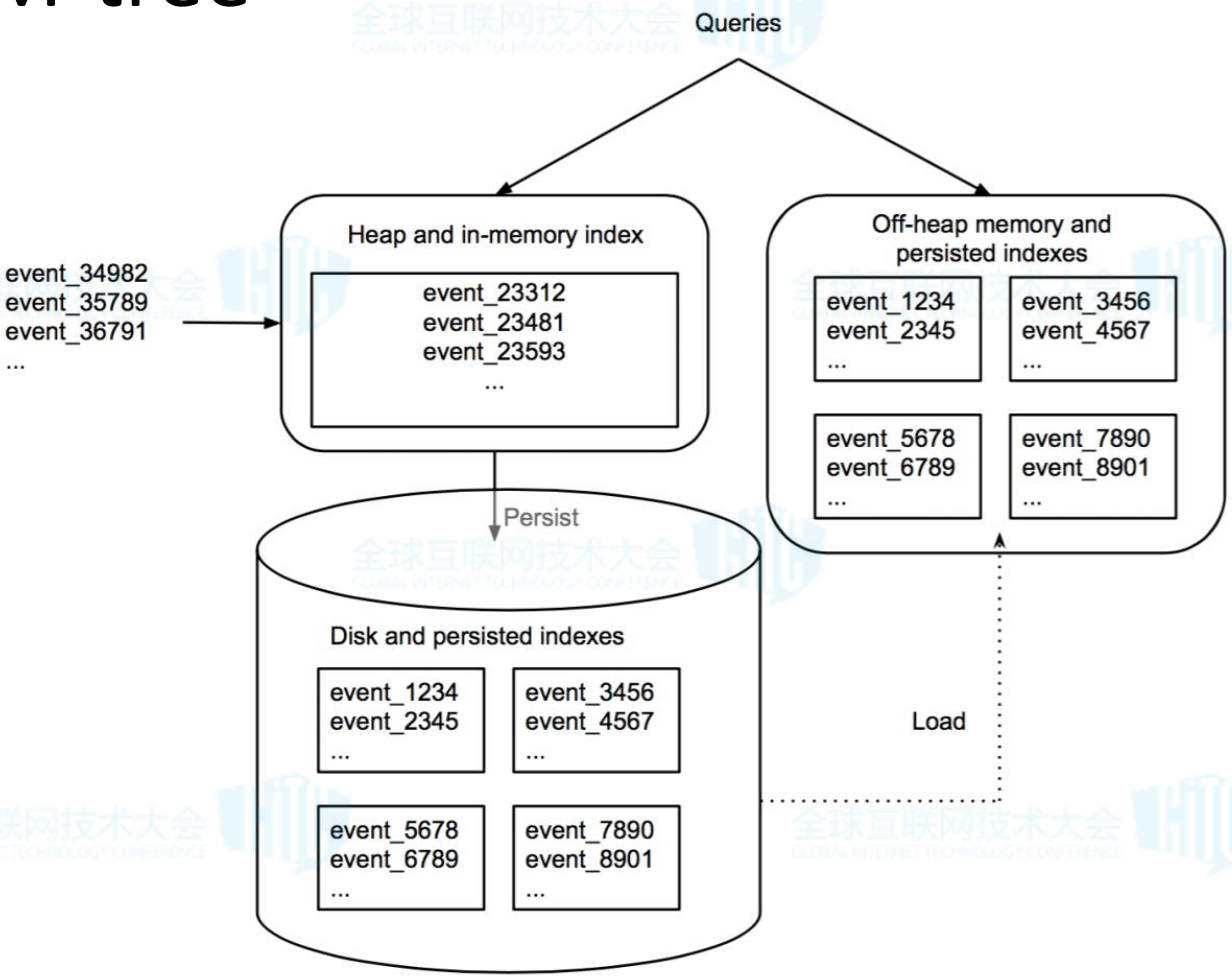


图 3-8　实时节点数据块的生成示意图

Druid一些高级特性

- 近似直方图和分位数
- 预估数据（Data Sketch）
- 地理索引和查询
- 路由器（Router）
- Kafka 索引服务

Druid数据分析生态系统

分析平台

- Imply (imply.io)
- Pulsar (eBay)

数据管理

- Druid-Spark-Batch (MetaMarkets)
- spark-druid-Olap (SparklineData)
- Calcite (Apache)

Druid管理

- Druid-Metrics-Kafka (quantiply)
- Druid-Console (druid.io)
- Druid-C (quantiply)
- Docker-Druid (druid.io)
- Tranqulity (druid.io)

Druid

数据可视化

- MetaBase
- Grafana-plugins (quantiply)
- Caravel (AirBnb)
- Pivot (imply.io)

访问扩展

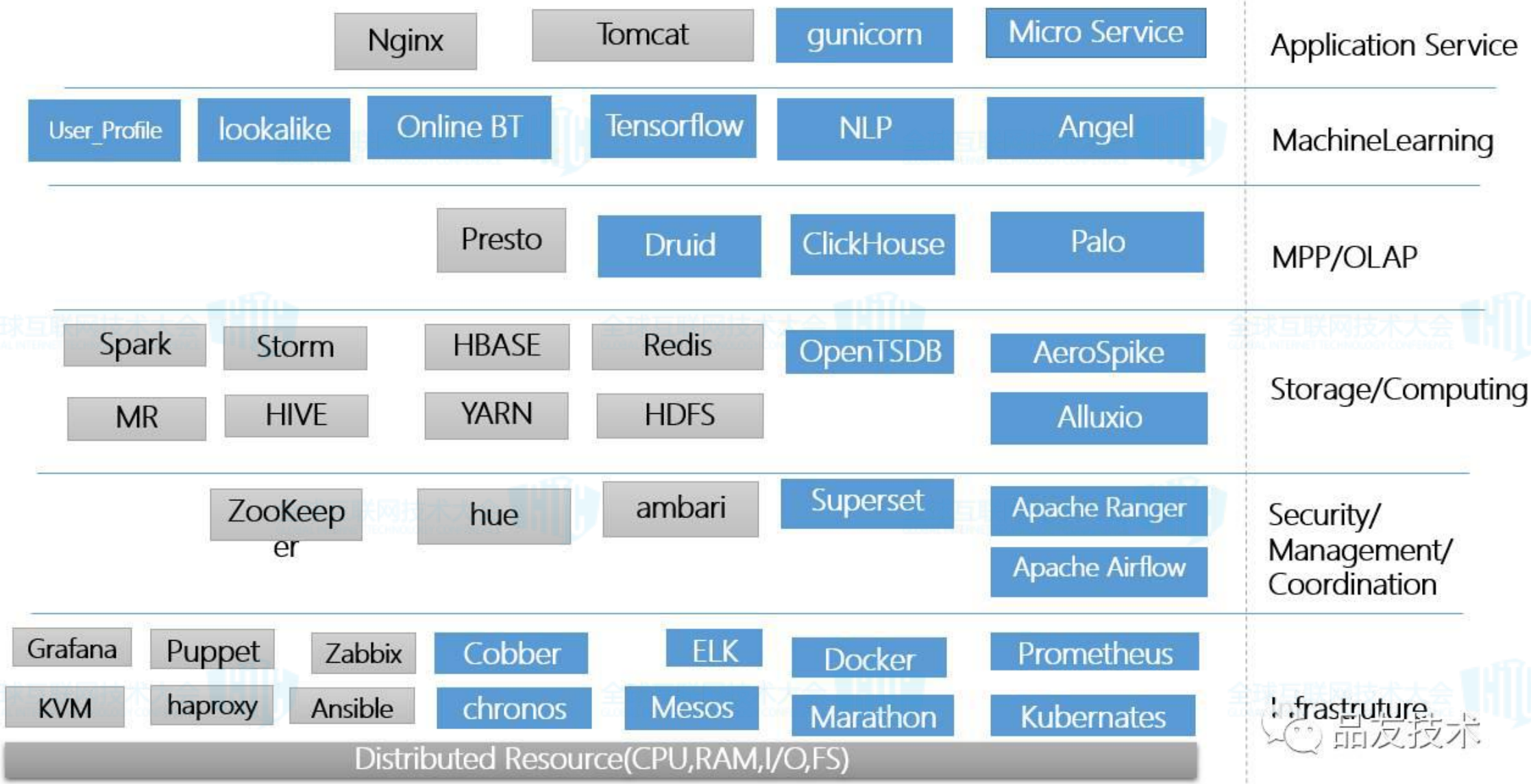
- PyDruid (druid.io)
- SQL4D (srikalyc)
- Calcite (Apache)
- RDruid (Druid.io)
- PlyQL (imply.io)

数据源

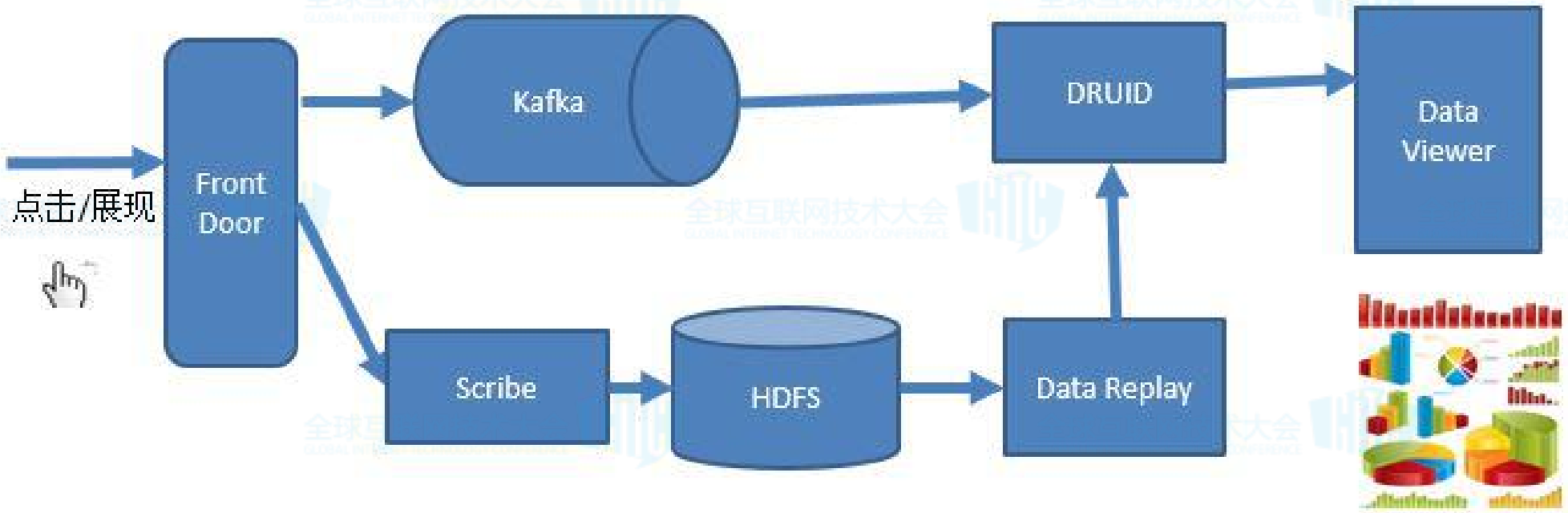
- Kafka
- HDFS
- Storm
- S3

	DRUID	Pinot	KYLIN
使用场景	实时处理分析	实时处理分析	OLAP分析引擎
开发语言	JAVA	JAVA	JAVA
接口协议	JSON	JSON	OLAP/JDBC
发布时间	2011	2015	2015
Sponsor	MetaMarkets /Yahoo	LinkedIn	eBay
技术	实时聚合	实时聚合	预处理，Cache

品友第三代数据分析平台的技术栈



Druid行业应用：程序化广告平台分析



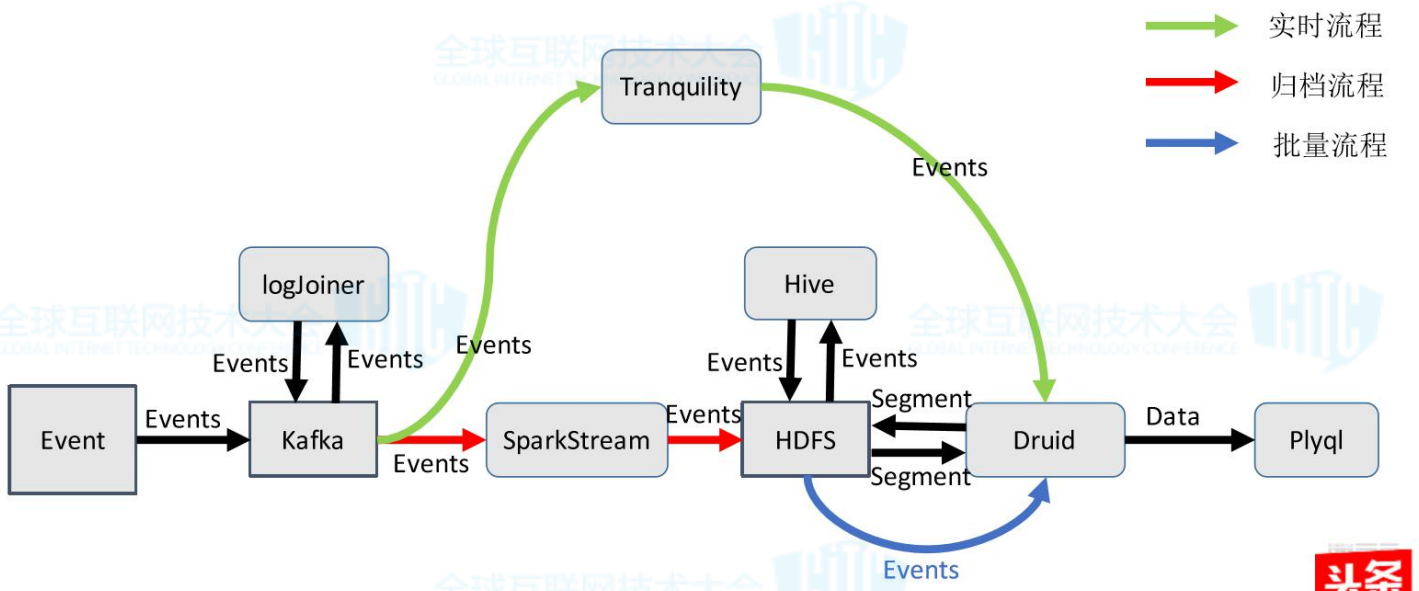
《Druid实时大数据分析》

Druid的应用：头条用户行为分析

应用情况

- 按查询最低粒度创建DataSource
  - 目前38个dimension，9个metric
  - 小时粒度的Segment，平均每个Shard 700MB，每天1.3T
- 按BI查询需求抽出中间表
  - 从低粒度表reindex出来，14个dimentsion， 9个metric
  - 天粒度的Segment，每个Shard 40MB，一个副本
- 按实时分析、计算与监控需求创建DataSource
  - 目前13个dimension，9个metric
  - 15分钟粒度的Segment，每个Shard 500MB，一个副本

系统架构



From:第四次中国Druid用户组Meetup

Druid的应用：OneAPM监控

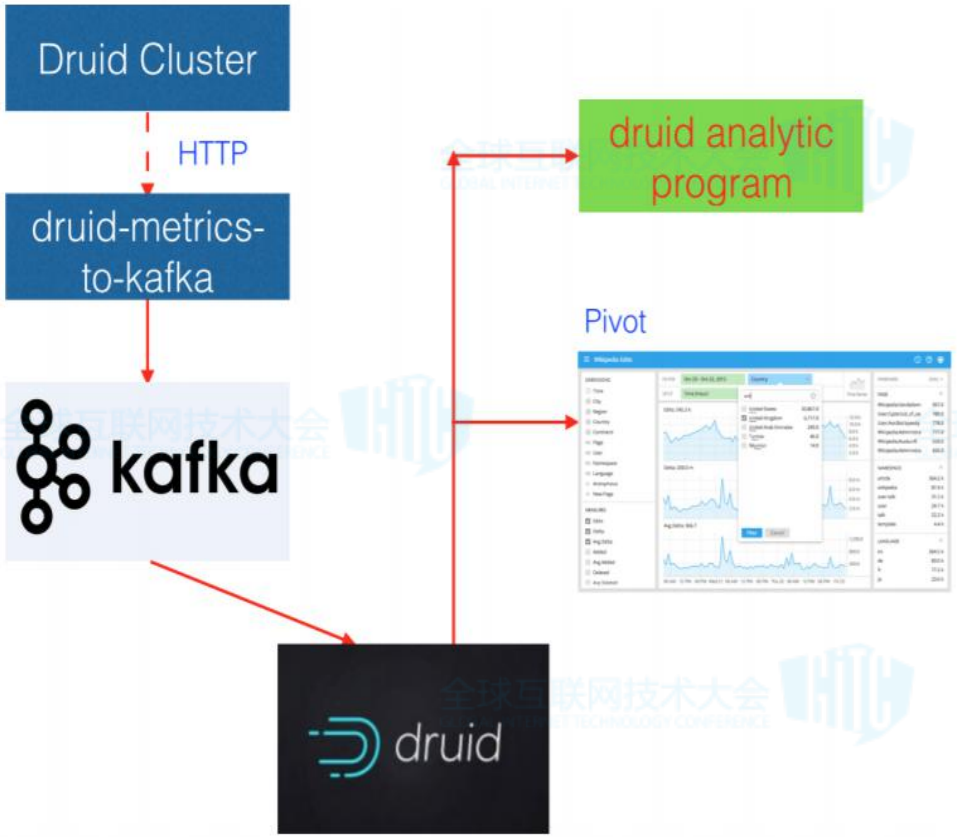


图 9-1　基于 HTTP 方法的监控系统架构示意图

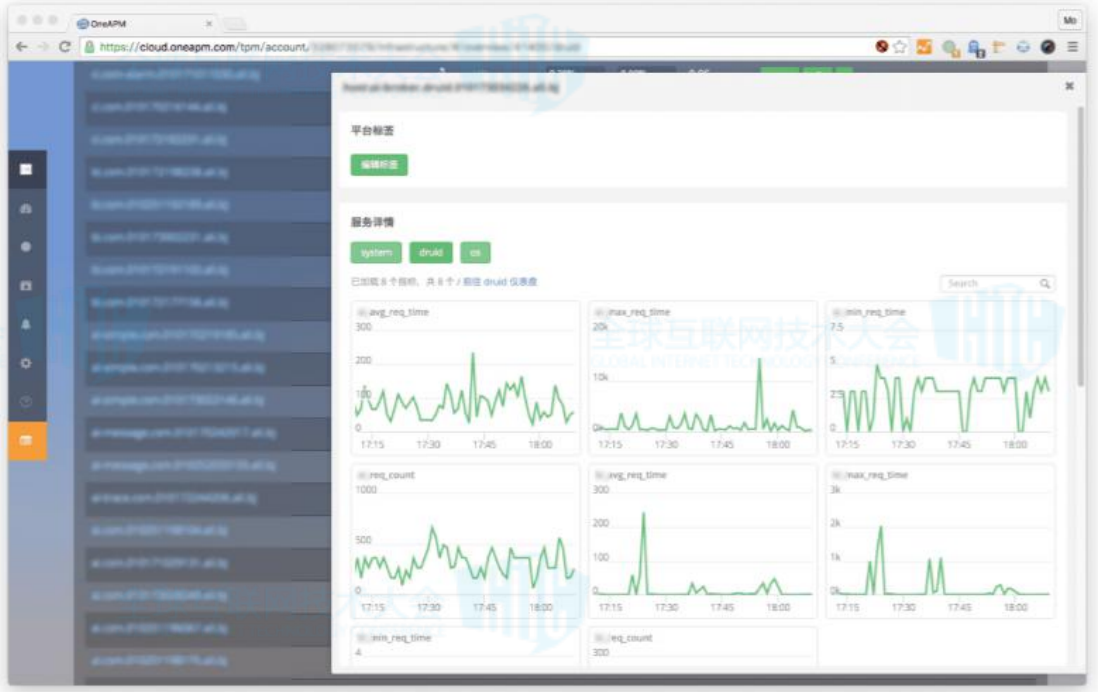
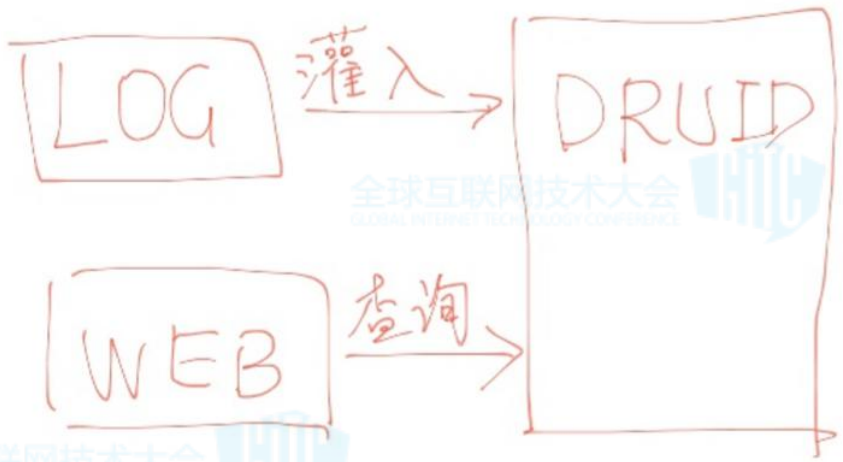


图 9-2　OneAPM Cloud Insight 对 Druid 监控的支持

《Druid实时大数据分析》

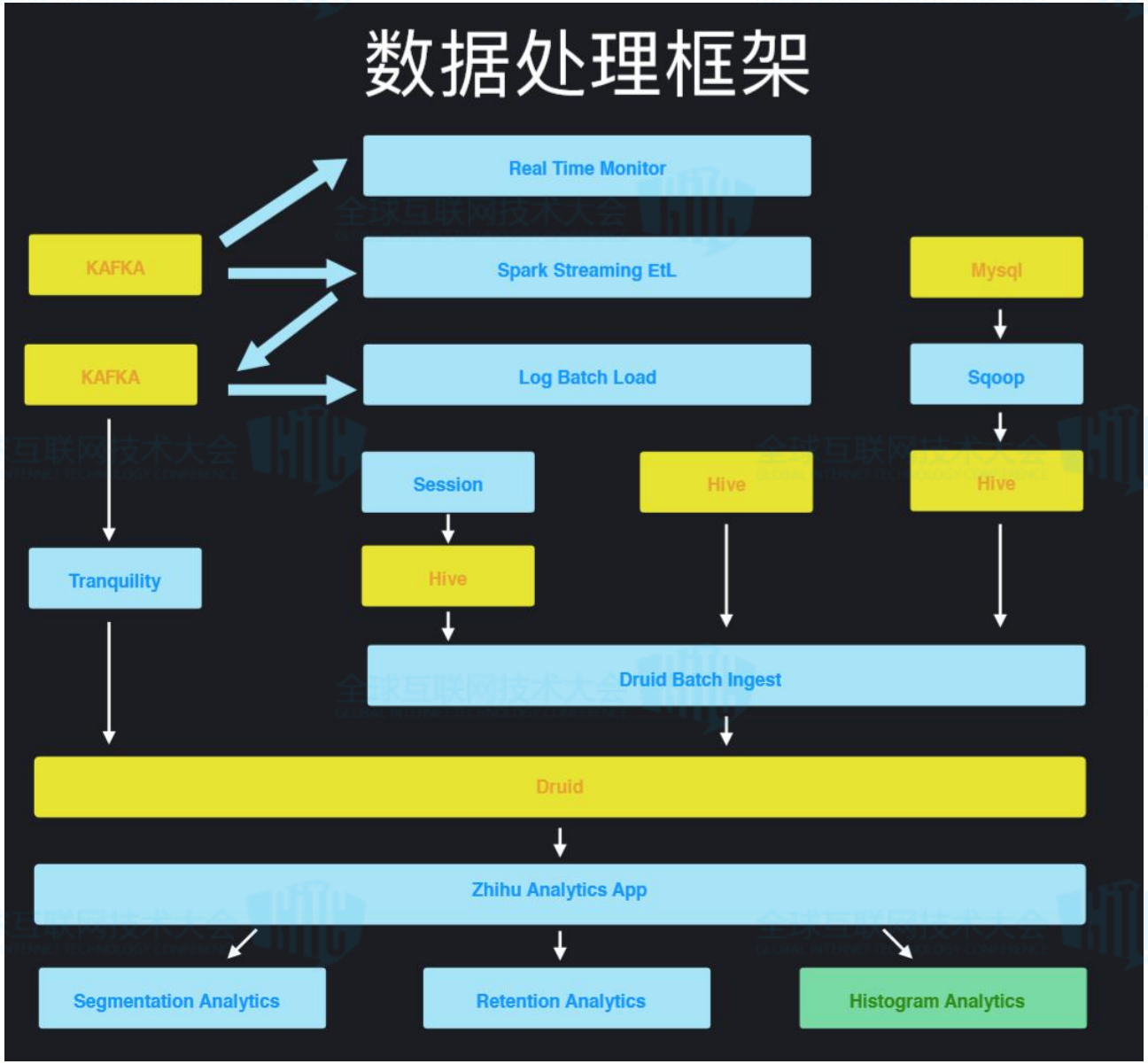
Druid的应用：网易行为分析

我们的需求	druid feature	备注
pv	count，longSum	
uv	datasketches	为了精度，size设置较大
新老用户	selector， filtered aggregator， post aggregator	数据在灌入druid前增加 isOldUser字段
回访	datasketches， filtered aggregator	数据在灌入前增加intDay 字段



From:第四次中国Druid用户组Meetup

Druid的应用：知乎数据分析



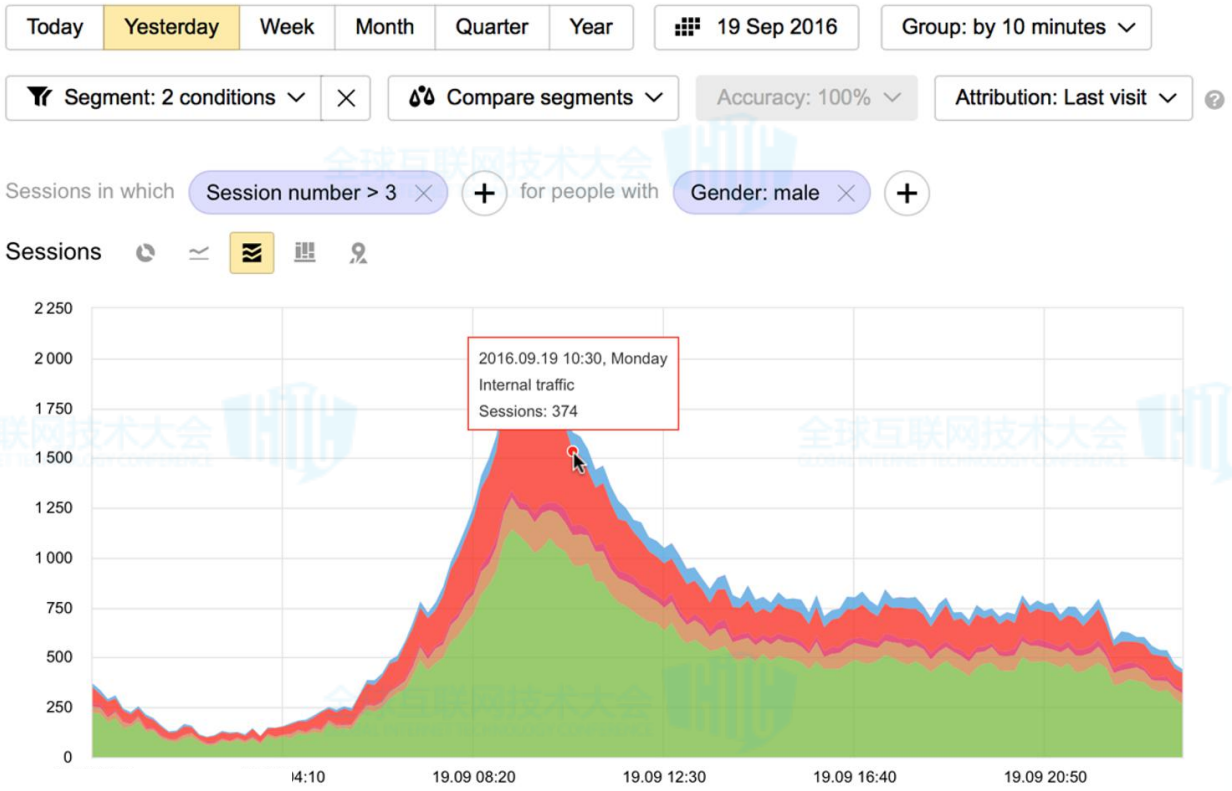
- 行为分析
- 留存分析
- 查询报表

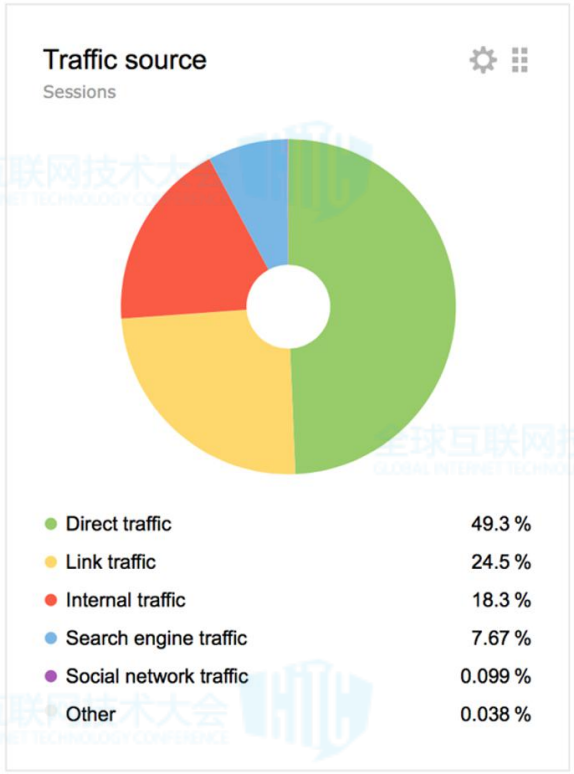
From:第四次中国Druid用户组Meetup

Druid不是银弹

- 虽然快，但是聚合是一个双刃剑
- 对CAP追求的执着
- 发现新物种？

Yandex.Metrica *(类百度统计 or Google Analytics)*





Yandex

Technology,
Technology,
Technology,
Local Focus,
Culture.

数据量：200+亿事件/天，100K+ 分析查询/天，数百万网站

ClickHouse的技术特性和不完美

ClickHouse 关键功能和应用场景

关键功能	应用场景
深度列存储	广告网络和RTB
向量化查询执行(Vectorized Query Execution)	电信
数据压缩	电子商务
并行和分布式查询	信息安全
实时数据注入	监测和遥感
跨数据中心的备份	商业智能
磁盘上的数据访问局部性（Locality of reference)	网络游戏
类SQL 支持	物联网
局部和分布式的Join	
可插入式的纬度表(dimension table)	
预估查询处理	
支持IPV6数据格式	
网站和应用分析	

https://clickhouse.yandex

ClickHouse的不完美：

1.不支持Transaction, OLTP

2.聚合结果必须小于一台机器的内存大小

3.缺少完整的Update/Delete操作

4.不适合典型的Key-Value存储

5.不支持Blob/Document类型数据

6.仅仅支持Ubuntu OS，其他用Docker

Event-oriented RDBMS

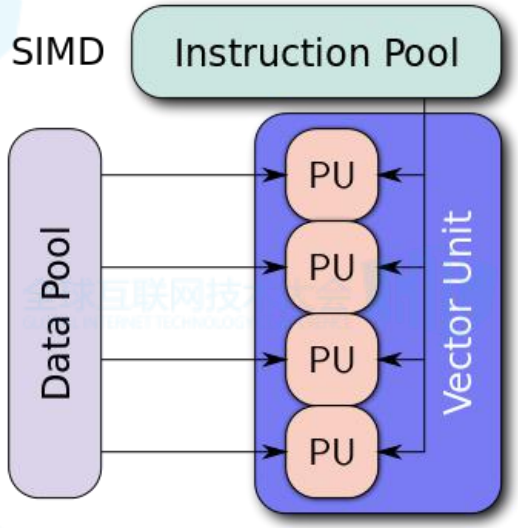
ClickHouse/Metrica发展简史

- 第一阶段MYISAM (LSM-Tree) (2008-2011)
- 阶段二: Metrage (从2010-现在/End)
- 阶段三: OLAPServer (2009-2013)
- 第四阶段:ClickHouse（2011-现在）

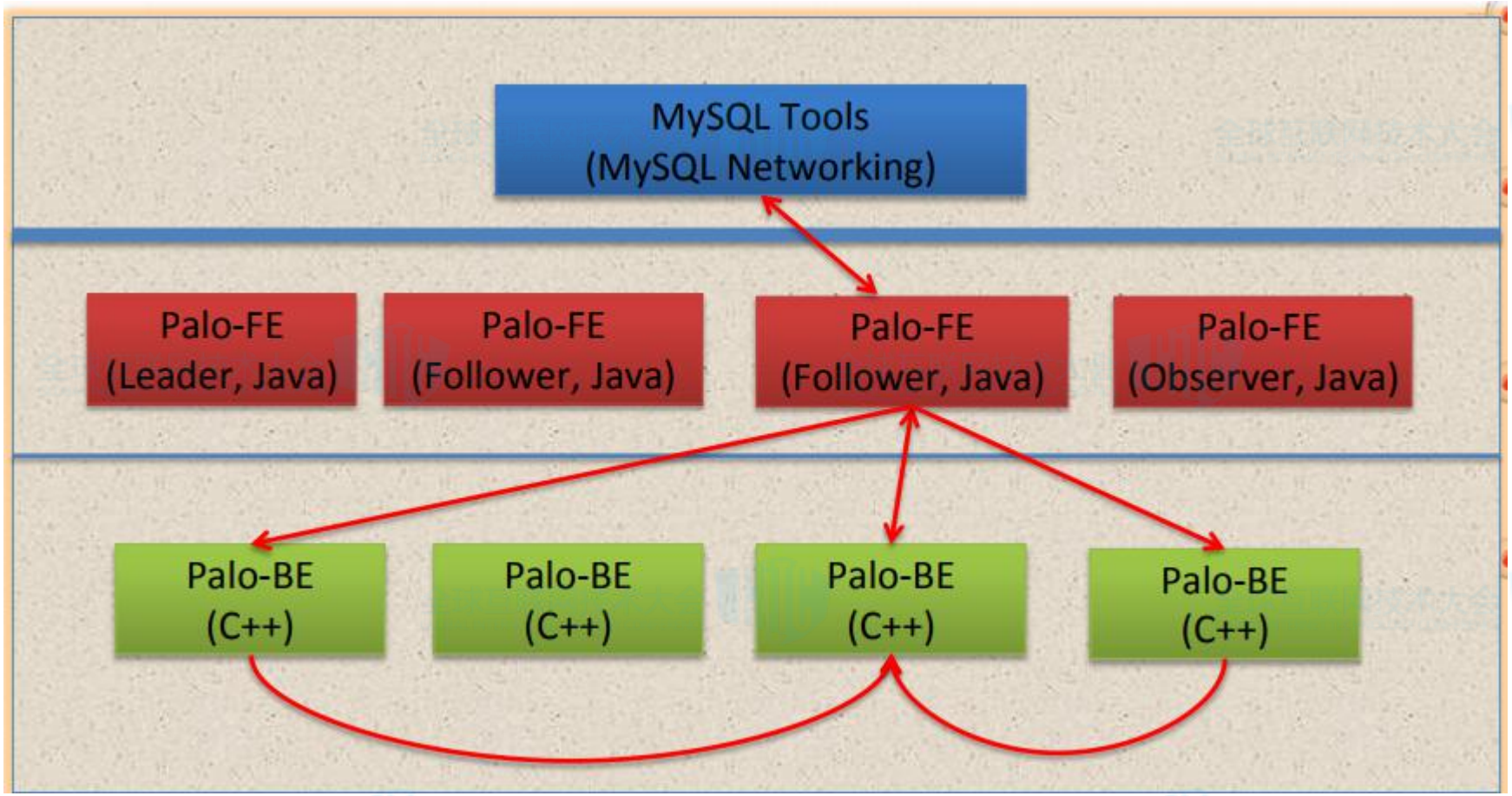
ClickHouse为什么这么快？

性能为王的原则，每一个改动都需要经过性能测试。

- Vectorized Query Execution技术
  - 利用CPU的SIMD （Single Instruction Multiple Data）
  - 来自VectorWise公司（Actian now!)
  - 参考“*Vectorization vs. Compilation in Query Execution*”
- Runtime Code Generation技术
  - Java JIT/Reflection; C++ LLVM；
- C++ 14特性
  - TCMalloc类似技术



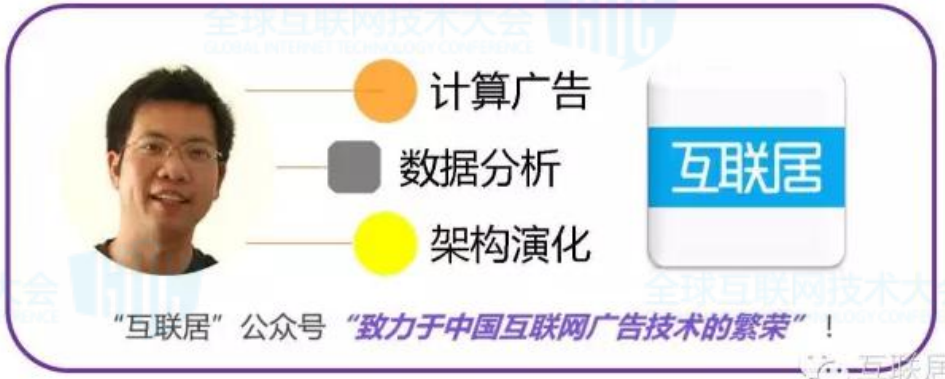
百度 Palo的整体架构

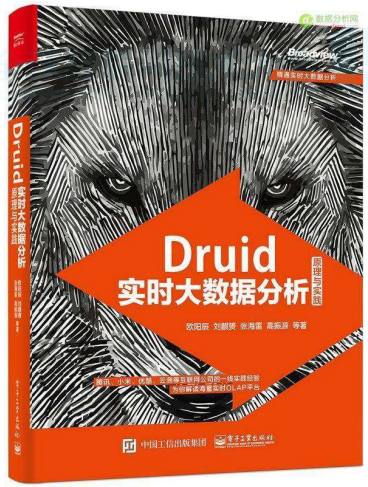


- 列存储和压缩
- 两层分区与分级存储 (SSD,SATA）
- 向量化&LLVM
- 物化视图 (切表）

谢谢！

互联居





www.ouyangchen.com