Geekbang 极客邦科技 InfoQ

QCon 全球软件开发大会【北京站】2016

百度大数据即席查询服务

Baidu BigSQL/孙垚光

Intemational Software Development Conference

自我介绍

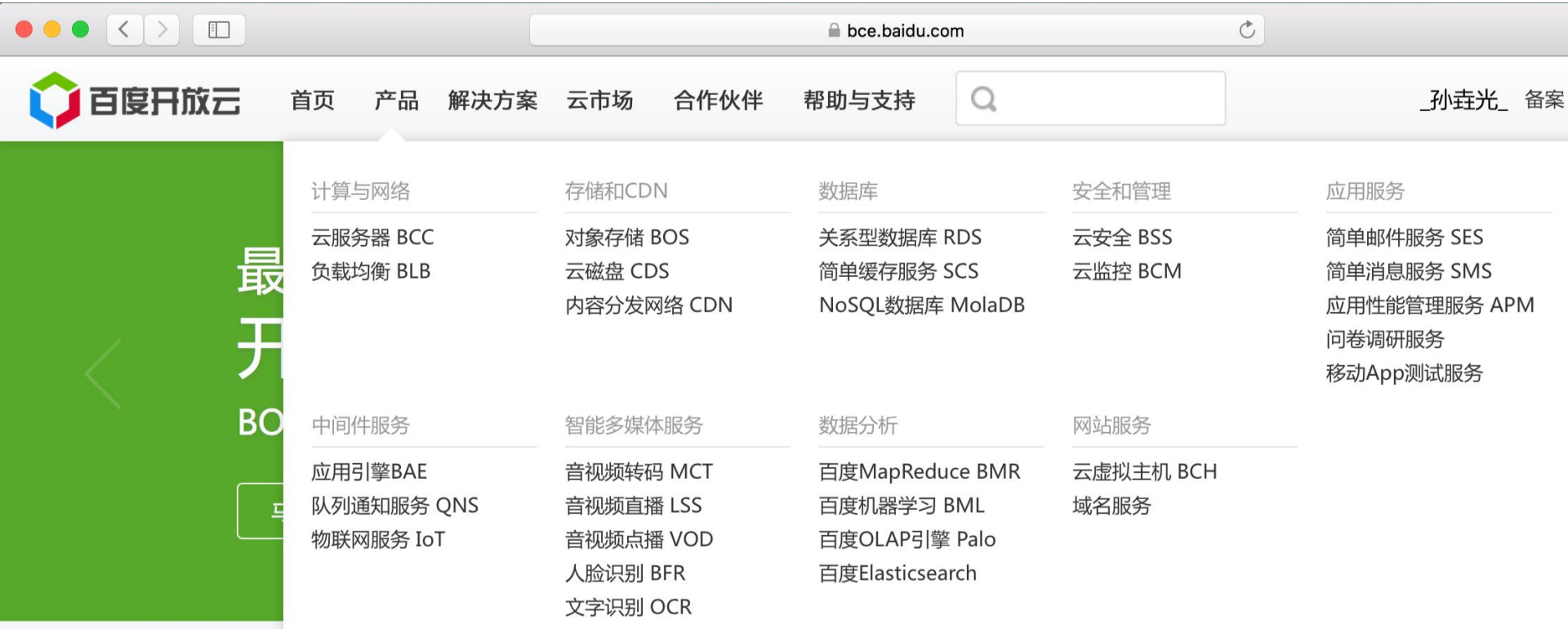
基础架构部 分布式计算团队 孙垚光

09年-11年：Linux内核/网络协议栈优化

11年-今：分布式计算/百度开放云

Hadoop/Spark

百度开放云


bce.baidu.com
百度开放云
首页
产品
解决方案
云市场
合作伙伴
帮助与支持
孙垚光
备案
计算与网络
云服务器 BCC
负载均衡 BLB
存储和CDN
对象存储 BOS
云磁盘 CDS
内容分发网络 CDN
数据库
关系型数据库 RDS
简单缓存服务 SCS
NoSQL数据库 MolaDB
安全和管理
云安全 BSS
云监控 BCM
应用服务
简单邮件服务 SES
简单消息服务 SMS
应用性能管理服务 APM
问卷调研服务
移动App测试服务
中间件服务
应用引擎BAE
队列通知服务 QNS
物联网服务 IoT
智能多媒体服务
音视频转码 MCT
音视频直播 LSS
音视频点播 VOD
人脸识别 BFR
文字识别 OCR
数据分析
百度MapReduce BMR
百度机器学习 BML
百度OLAP引擎 Palo
百度Elasticsearch
网站服务
云虚拟主机 BCH
域名服务

即席查询服务（BigSQL）

- BigSQL定位/特点
- BigSQL架构
- BigSQL关键技术
- BigSQL在Baidu内部的应用
- 下一步计划

即席查询服务（BigSQL）

- **BigSQL定位/特点**
- BigSQL架构
- BigSQL关键技术
- BigSQL在Baidu内部的应用
- 下一步计划

BigSQL 定位

- 大数据即席查询（Ad-Hoc Query）平台
- PAAS： 开箱即用，用户无需关心机器/集群的运维/细节
- 高性能/规模： 裸机/优化/最大PB量级以上
- 低成本： 多租户共享集群/按使用付费

BigSQL 特点

- 数据格式：半结构化（CSV/JSON/Parquet/Protobuf等）
- 使用接口：易用/多样化（RestAPI/Console/CLI/JDBC）
- 语法集：兼容开源SparkSQL/HQL
- 按使用付费：按（Query复杂度+扫描数据量）计费
- 多用户协同：灵活的权限管理

Ad-Hoc Query

- 面向“人”的查询
  - ✓交互式（Interactive）：
    - ➢查询具有较高时效性
  - ✓即席（Ad-Hoc）：
    - ➢查询模式相对不固定
    - ➢数据没有（时间/成本）做过多预处理

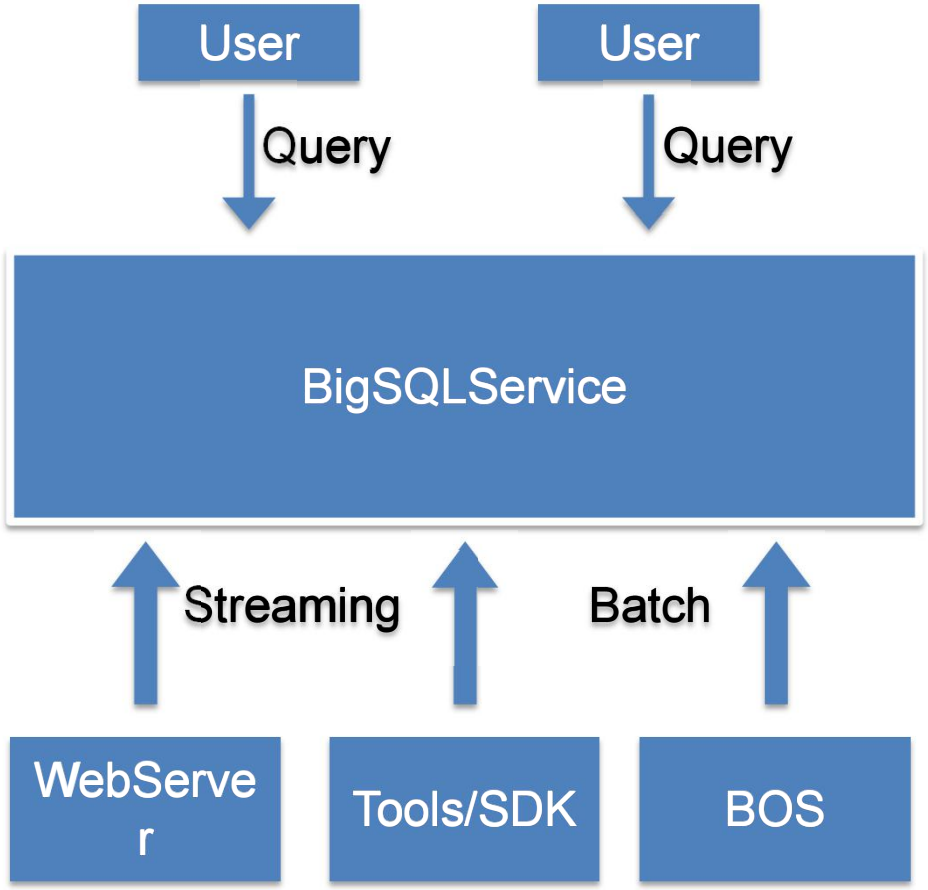
即席查询 vs 多维分析

	Ad-Hoc Query	OLAP
数据密度	弱（半）结构化	高度结构化
加工过程	粗（浅）加工	深度加工
查询模式	相对随机	相对固定

MPP/Shared-Nothing

	MPP/Impala	SQL on Hadoop/SparkSQL
扩展性	1000台以内/PB以下	千台以上/PB以上
查询延迟	毫秒~秒	秒~分钟
架构复杂性	中等	复杂
容错	无	有
调度策略	Gang/Transaction	分批
启停开销	小/常驻进程	大/现启动
与存储结合程度	紧密	松散

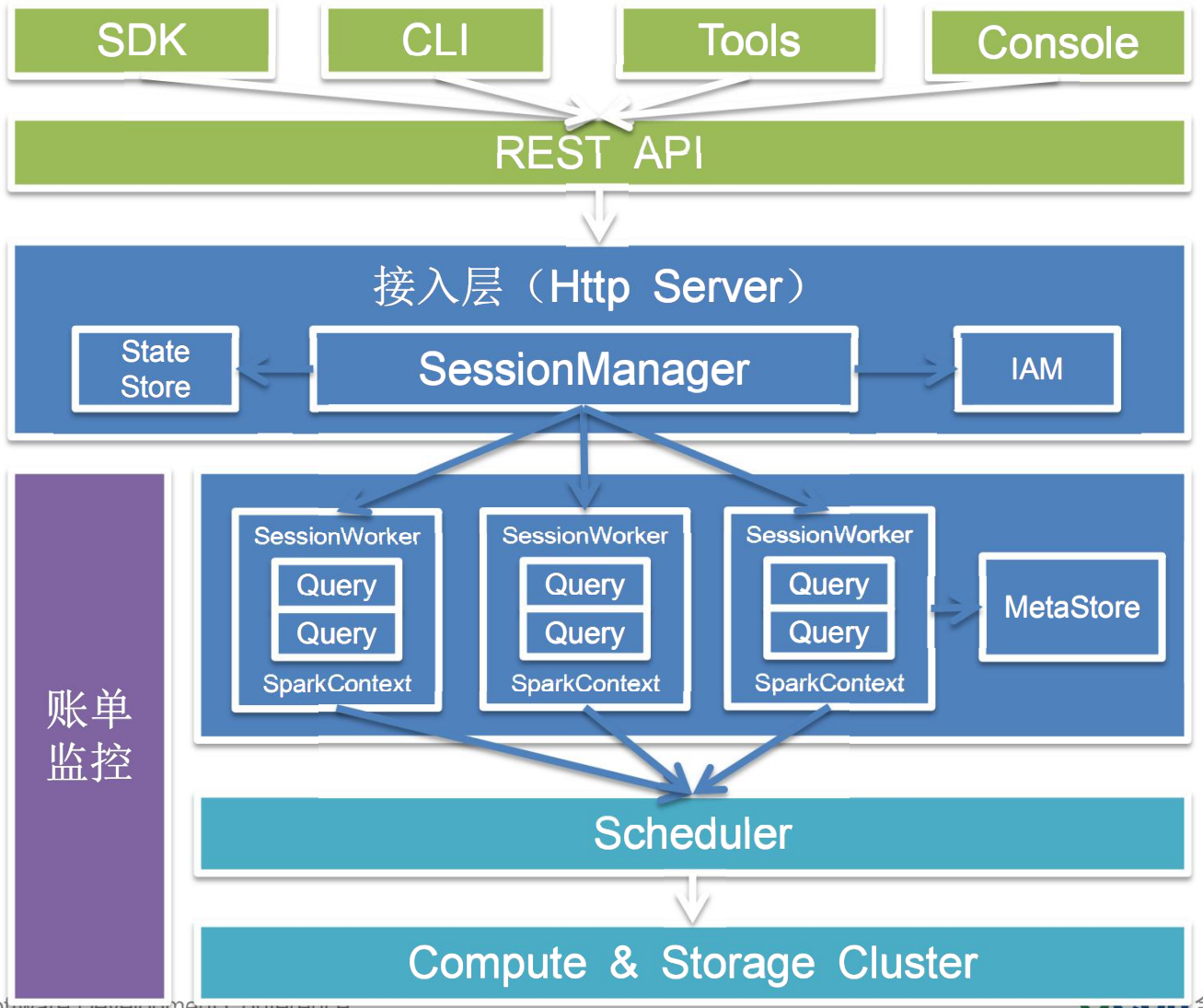
BigSQL 示意图


User
User
Query
Query
BigSQLService
Streaming
Batch
WebServe
r
Tools/SDK
BOS

即席查询服务（BigSQL）

- BigSQL定位/特点
- **BigSQL架构**
- BigSQL关键技术
- BigSQL在Baidu内部的应用
- 下一步计划

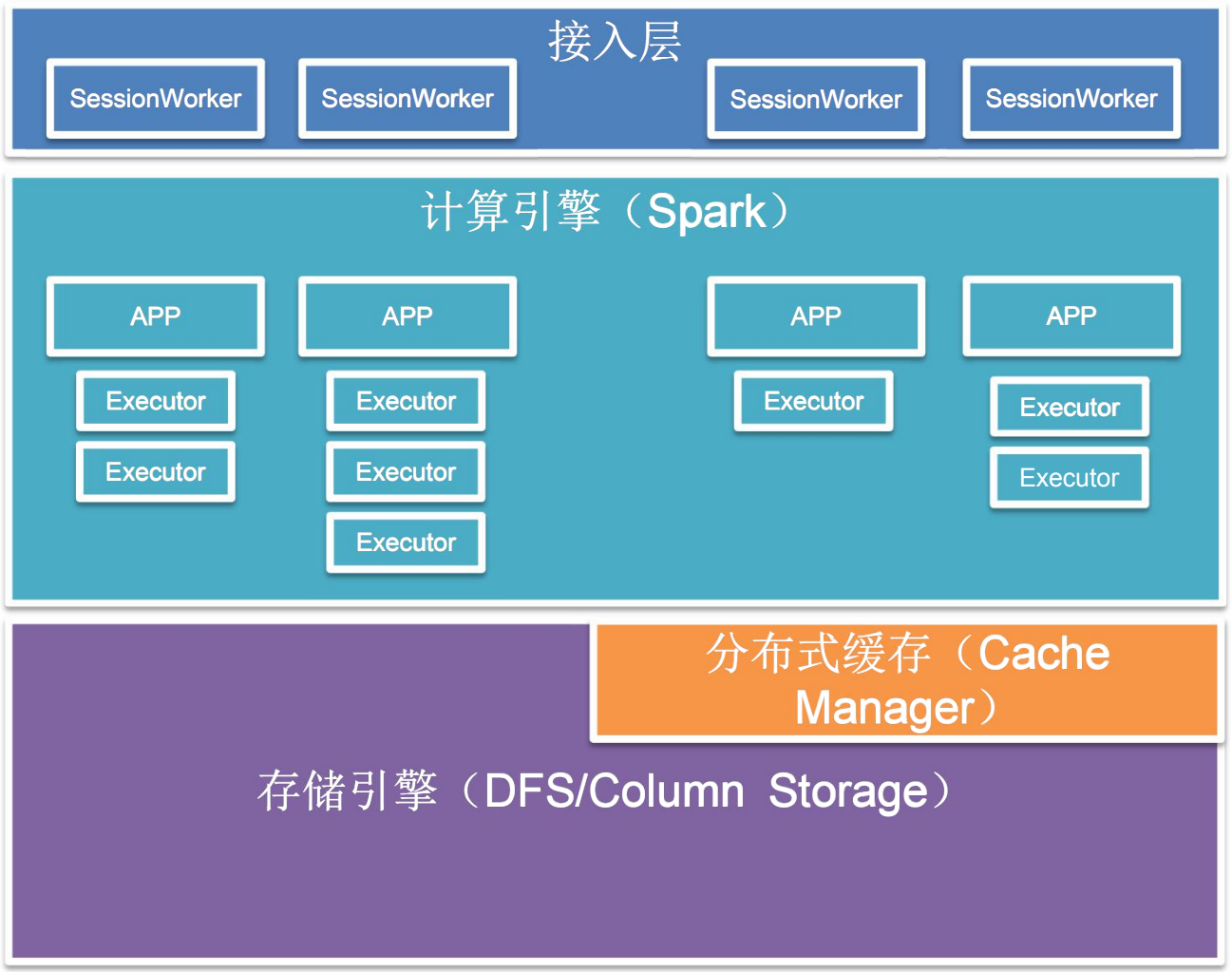
BigSQL 整体架构


SDK
CLI
Tools
Console
REST API
接入层（Http Server）
State Store
SessionManager
IAM
账单
监控
SessionWorker
Query
Query
SparkContext
SessionWorker
Query
Query
SparkContext
SessionWorker
Query
Query
SparkContext
MetaStore
Scheduler
Compute & Storage Cluster

BigSQL整体架构：接入层

- 易用性：各种形式的API
- 可用性：关键节点容错
- 安全：租户认证和鉴权、Quota限制
- 账单
- 监控

BigSQL整体架构：引擎层

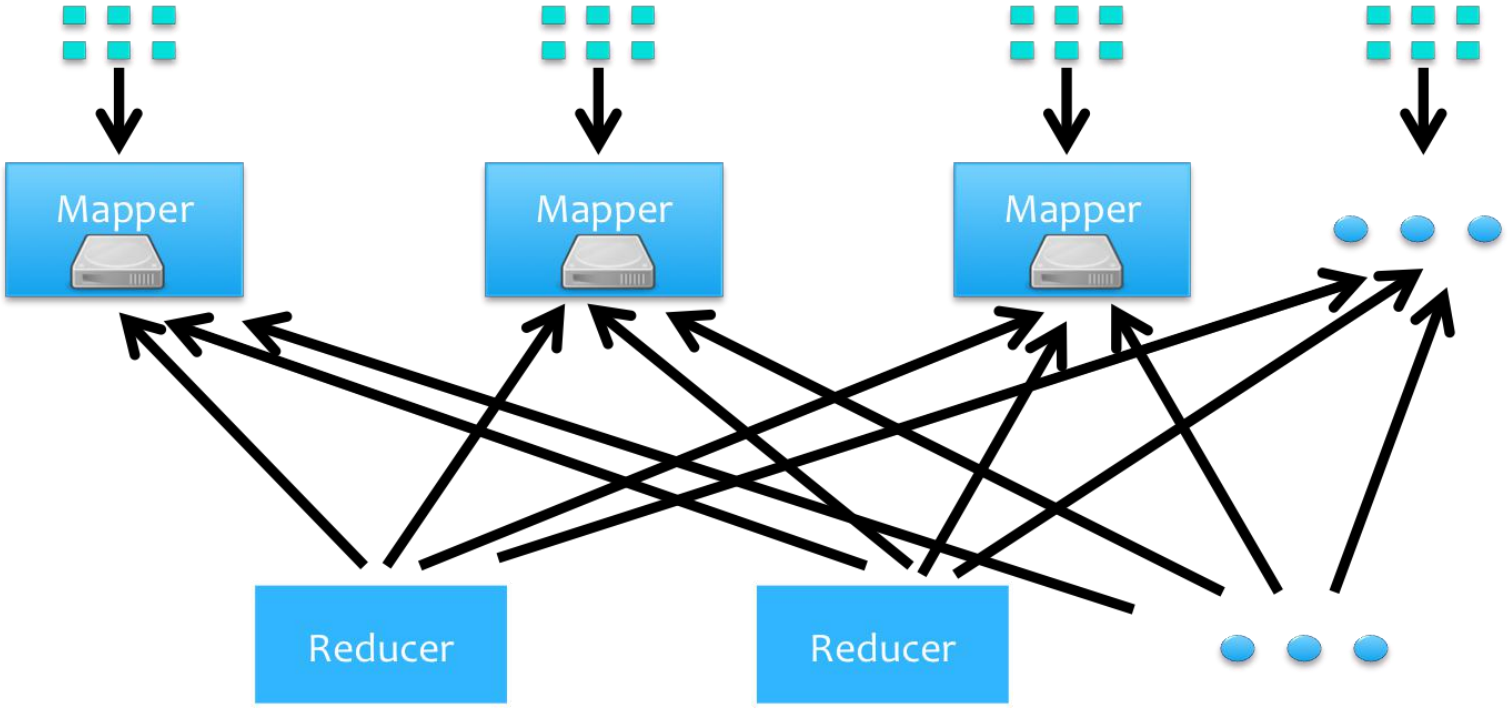


即席查询服务（BigSQL）

- BigSQL定位/特点
- BigSQL架构
- **BigSQL关键技术**
- BigSQL在Baidu内部的应用
- 下一步计划

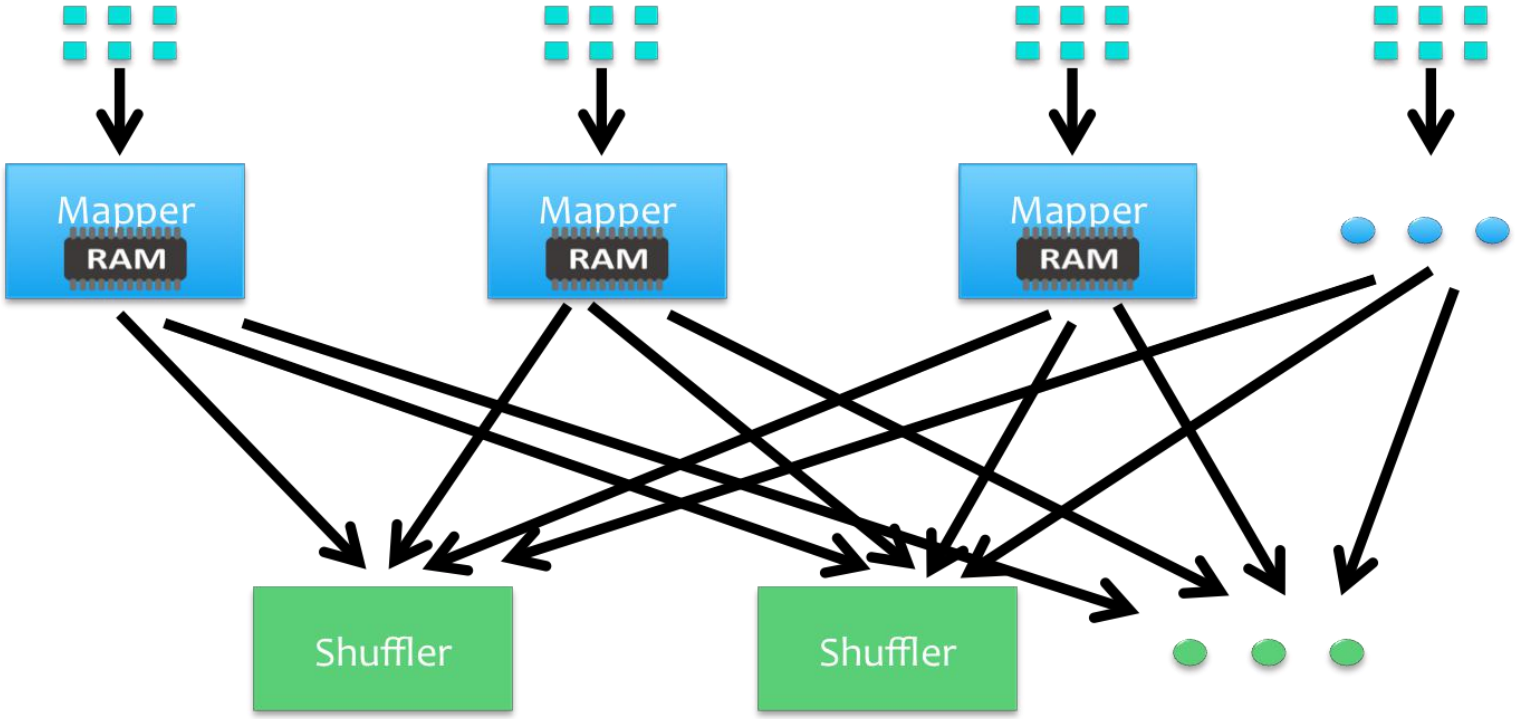
BigSQL 关键技术（一）

高性能Shuffle



BigSQL 关键技术（一）

高性能Shuffle



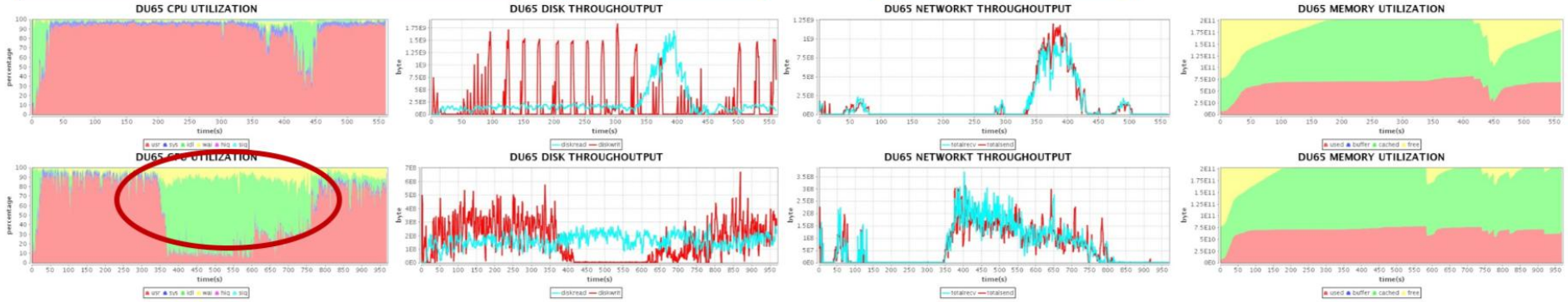
BigSQL 关键技术（一）

SSD v.s. HDDs: x1.7 end-to-end improvement

x1.7 end-to-end improvement
x5 shuffle improvement

x6 Disk BW

x4 network BW

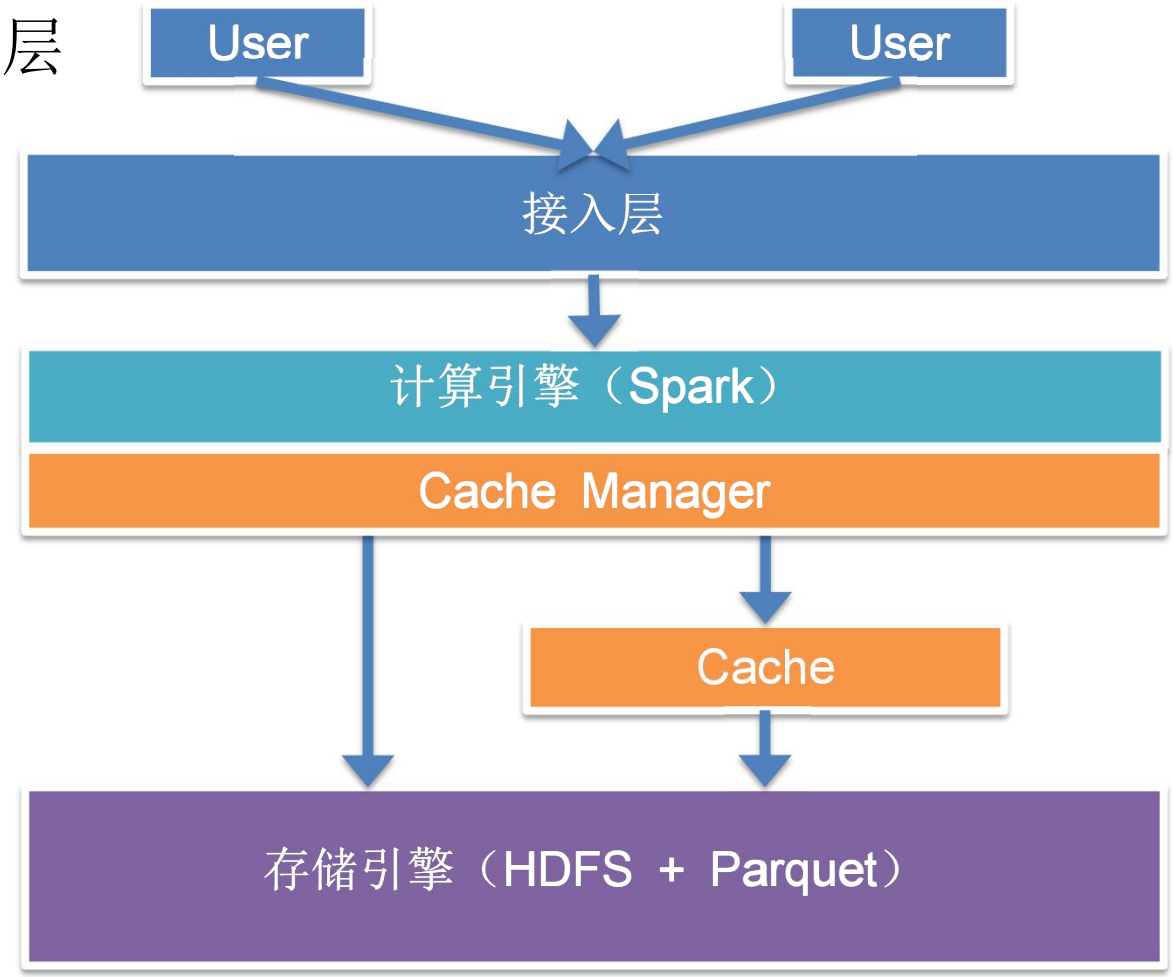


Device:	rrqm/s	wrqm/s	r/s	w/s	rkB/s	wkB/s	avgrq-sz	avgqu-sz	await	svctm	%util
nvme0n1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
sda	0.00	0.00	0.00	9.00	0.00	36.00	8.00	0.00	0.00	0.00	0.00
sdc	2.00	0.00	398.00	0.00	37268.00	0.00	187.28	7.09	15.63	2.46	97.90
sdb	19.00	347.00	389.00	8.00	35236.00	1420.00	184.66	10.37	25.28	2.45	97.30
sdd	19.00	0.00	326.00	23.00	29592.00	11776.00	237.07	153.35	137.34	2.87	100.00
sde	0.00	0.00	317.00	0.00	30344.00	0.00	191.44	2.87	9.02	2.54	80.40
sdf	11.00	0.00	267.00	1.00	25332.00	4.00	189.07	12.98	50.60	3.73	100.00
sdg	18.00	332.00	384.00	0.00	34684.00	0.00	180.65	25.56	47.58	2.60	100.00
sdh	4.00	183.00	334.00	5.00	33392.00	752.00	201.44	4.86	14.35	2.69	91.30

Significant IO bottleneck

BigSQL 关键技术（二）

数据缓存层



BigSQL 关键技术（二）

数据缓存策略

- 按需缓存
  - Query运行时触发Cache miss，异步load到缓存
- 数据预取
  - 周期性Load相关Table/Partition到缓存
  - 根据过去Query信息统计热点数据，提前Load到缓存

典型案例：跨地域查询加速（提升至少一个数量级）

BigSQL 关键技术（三）

优化执行

- 智能参数优化
  - 利用Combine类InputFormat，减少MapTask数
  - 根据上游输出，自动优化Reduce Partition数目
- 调度优化
  - 评估数据量，自动复用Application 或者 启用新的Application
- 近似查询
  - 长尾任务自动忽略，保证时效性

BigSQL 关键技术（四）

资源隔离/安全

- 基于Cgroup/Namespace的Container隔离
  - CPU/Memory/FS
  - Container本身的加固
  - 网络的互通与隔离
- JVM沙箱层的多种安全策略
- 计算/存储框架层的安全认证和加密传输

即席查询服务（BigSQL）

- BigSQL定位/特点
- BigSQL架构
- BigSQL关键技术
- **BigSQL在Baidu内部的应用**
- 下一步计划

在Baidu内部的应用

凤巢广告数据分析

- 漏斗分析
  - 分析广告被过滤的原因，各个维度特征等
- 系统优化和问题定位
  - 分析系统业务日志，发现可优化的指标和潜在问题

日均扫描数据量：xx PB

即席查询服务（BigSQL）

- BigSQL定位/特点
- BigSQL架构
- BigSQL关键技术
- BigSQL在Baidu内部的应用
- 下一步计划

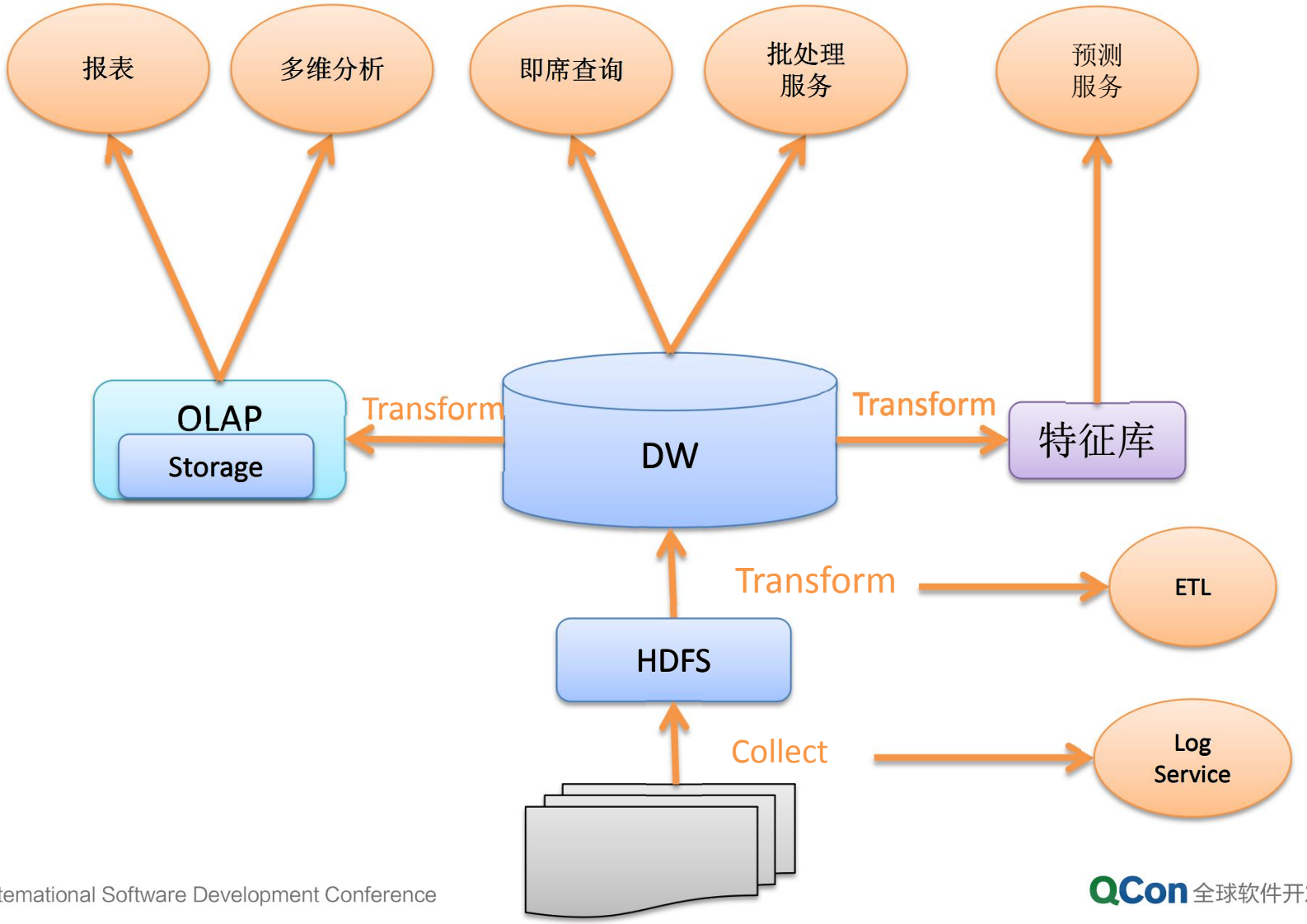
BigSQL 后续规划

- 持续投入技术研发
  - 更智能的数据缓存层： 细粒度/物化视图选取
  - 实时更新
  - 向量执行： 提高CPU cache命中率
  - CBO： Cost-based Optimizer

BigSQL 后续规划

- 构建通用大数据处理平台
  - 日志收集服务
  - 数据变形/ETL服务
  - 报表/多维分析
  - 即席查询服务
  - 批处理服务
  - 预测服务

通用大数据处理平台


报表
多维分析
即席查询
批处理
服务
预测
服务
OLAP
Storage
Transform
DW
Transform
特征库
Transform
ETL
HDFS
Collect
Log
Service

Geekbang 极客邦科技 InfoQ

THANKS!