

DTCC

2013中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2013
大数据 数据库架构与优化 数据治理与分析

SequeMedia
盛拓传媒

IT168.com

ITPUB

ChinaUnix

云端漫步

使用AWS搭建高伸缩数据 处理平台

Database
BDaaS
flowingdata
DB2
NoSQL MySQL
Oracle Big Data

兰亭集势 - 整合跨国供应链

Lightinthebox.com Shopping Global One World One Price

United States | \$ USD | English | Français | Español | Deutsch | More

TODAY'S SALES | DEALS | Sign In or Register | My Account | Help | Live Chat

SHOP ALL CATEGORIES

- Weddings & Events (22369)
- Home & Garden (21526)
- Fashion (28935)
- Electronics (8473)
- Mobile Phones & Tabl... (521)
- Health & Beauty (10089)
- Toys & Hobbies (11576)
- Sports & Outdoors (6965)
- Gadgets & Accessorie... (6589)
- Clearance (4035)

Deal of The Day

CLARINS™

USD \$65.00
Value: USD \$95.40

Shop Now

Sprinkle LightinTheBox

Be Stylish
Be Different
Up To 85% Off
Shipping
April 16 - 22
SHOP NOW

Dresses Designed Just for You
Up To 80% Off
Shop Now

Weddings & Events

- Wedding Dresses
- Wedding Party Dresses
- Special Occasion Dresses
- Accessories
- Wraps

Special Occasion Dresses

- Weekly Deal
- Little Black Dresses
- Sequined Dresses
- Bandage Dresses
- Prom 2013

Fashion

- TS® Clothing
- Women's Clothing
- Fashion Lingerie
- Mens Fashion Clothing
- Fashion Handbags

Health and Beauty

- Tattoos, Body Art
- Wigs
- Makeup Supply
- Nail Care and Polish
- Health Care

兰亭集势 - 整合跨国供应链

- 每天100多个国家的数百万访客
- 每天数万多个包裹发往 100多个国家和地区
- 19种语言提供服务
- 数十万种在售商品

兰亭集势 - 整合跨国供应链

- 每天新增数据100G
- 每天数据计算吞吐量5T

兰亭数据处理的特点

从数据特点上看

- ✓ 离线分析，实时服务
- ✓ 原始信息非结构化
- ✓ 数据产生于国外

兰亭数据处理的特点

从数据来源上看

- ✓ 来自于 Web Server 的 Clickstream
- ✓ 来自于定制 Collect Server 的各种 Tracking 信息

兰亭数据处理的特点

从计算的需求上看

- ✓ 集群规模不大
- ✓ 集群使用不饱和
- ✓ 需要较大的计算弹性

兰亭数据分析的类型

从分析角度来看

- ✓ 用户行为分析
- ✓ 品类效能分析

兰亭数据分析的类型

从用途来看

- ✓ 运营指标分析
- ✓ 广告效果评估与自动竞价
- ✓ 商品陈列优化
- ✓ 帮助购买（商品关联推荐与促销）
- ✓ A/B Test结果评估

广告效果评估和自动化竞价

- 渠道效果评估
- Adwords自动竞价

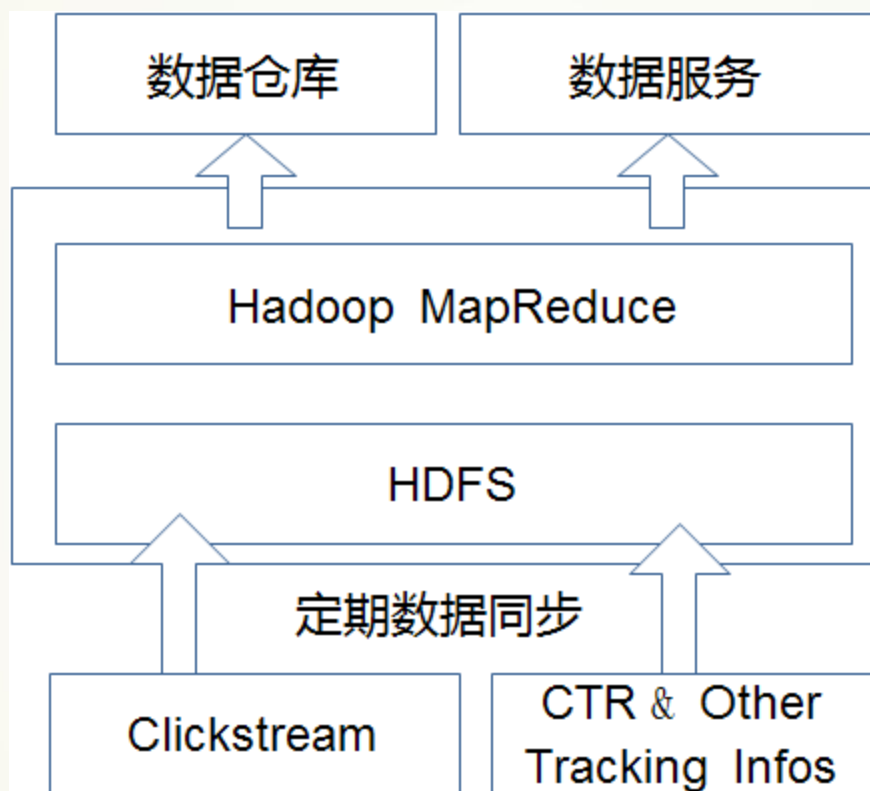
商品陈列优化

- 商品效能评分，自动晋级、淘汰
- 新品曝光

关联推荐与个性化导购

- 关联推荐
- 个性化导购

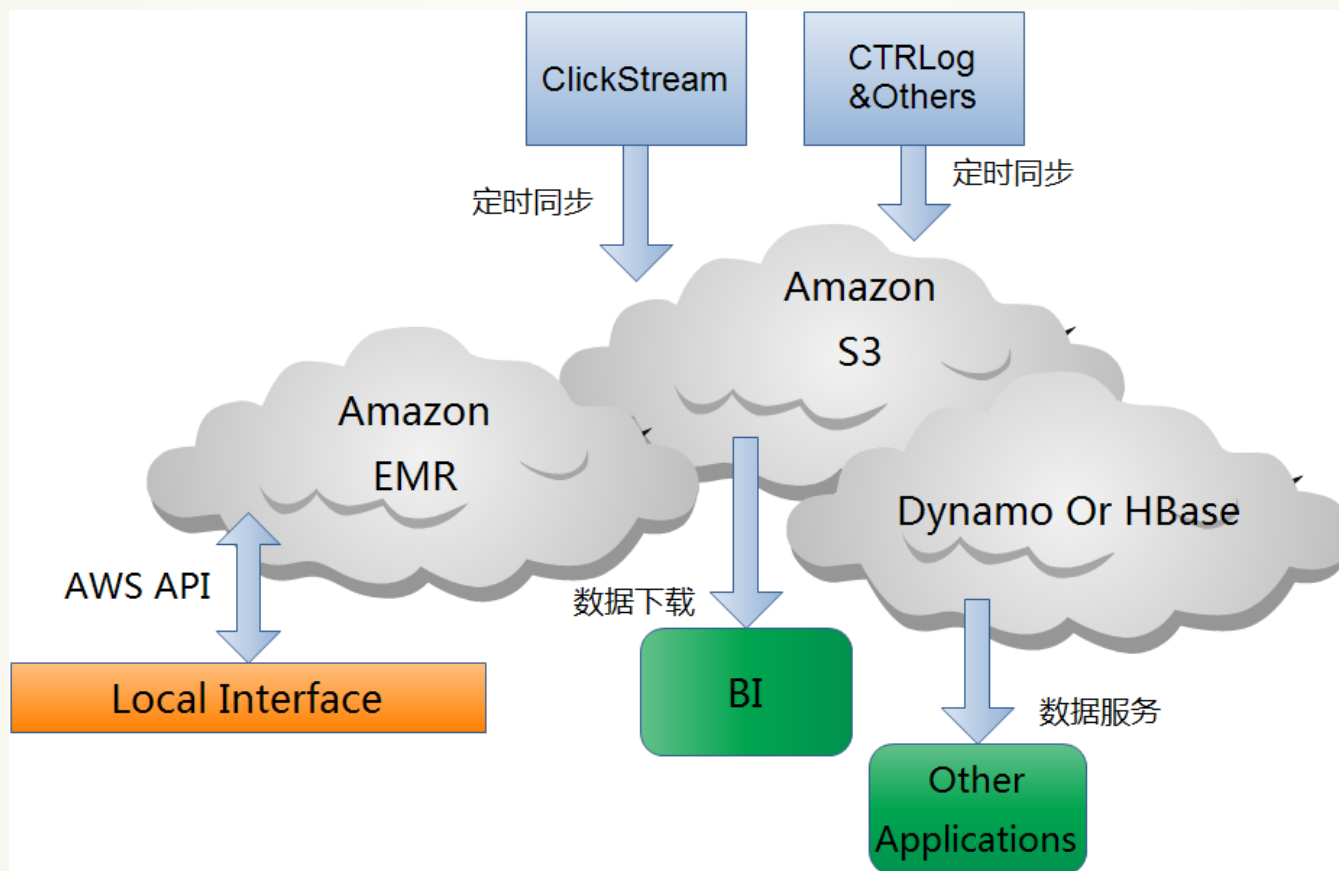
原有数据处理平台架构



原有数据处理平台的不足

- 数据与距离远，跨国传输成本高
- 存储能力与计算能力是静态的，计算与存储绑定，受集群节点数目限制，必须通过增加硬件投入来提高存储和计算能力
- 计算能力浪费与计算能力不足并存，缺乏弹性

使用AWS的数据处理平台架构



使用到的AWS服务

EC2 - Elastic Compute Cloud

Amazon 提供的基础计算虚拟实例

S3 - Simple Storage Service

Amazon 提供的存储服务

EMR - Elastic MapReduce

运行在EC2上的 MapReduce 环境

Amazon S3 – Simple Storage Service

- 几乎无限的数据存储
- 根据地理分布的存储位置
- 完善的安全管理
- 接口操作简便

Amazon EC2 - Elastic Compute Cloud

<http://aws.amazon.com/ec2/>

多种类型实例

高性能计算实例、高I/O实例、高存储实例

多种租用方式

按需租用

预留租用

Amazon EC2 - Elastic Compute Cloud

丰富的软件市场 (MarketPlace)

预置软件，自动部署，一键启动

各种辅助服务

- ✓ CloudWatch
- ✓ Auto Scaling

Amazon EMR – Elastic MapReduce

<http://aws.amazon.com/elasticmapreduce>

运行在在 EC2 上的 Hadoop 环境

支持Hive \ Pig \ Hbase



如何使用AWS

Web Management Console

<http://console.aws.amazon.com>

Web Services API

Custom Clients

两种平台架构的对比

本地 Hadoop

优点：调试方便，运行结果可及时检查

缺点：受集群硬件限制，计算能力不能及时调整

Elastic MapReduce

优点：计算能力按需调整

缺点：结果保存在S3上，调试不便

为什么要使用 EMR 替代本地 Hadoop

- ✓ 降低数据传输成本
- ✓ 增加数据存储能力扩展性
- ✓ 增强计算能力的扩展性
- ✓ 提高伸缩响应速度
- ✓ 摊薄存储与计算成本

开发自定义本地接口调用AWS

- ✓ 降低学习曲线，简化使用
- ✓ 透明化AWS与本地集群的调用切换
- ✓ 方便其他程序模块调用
- ✓ 加入权限管理与费用控制
- ✓ 帮助提高代码质量，减少成本浪费

自定义接口的安全控制

- ✓ 账号集中管理
- ✓ 账号权限与本地管理策略一致
- ✓ 细粒度控制不同账号在S3上的操作权限
- ✓ 集群规模控制
- ✓ 费用控制

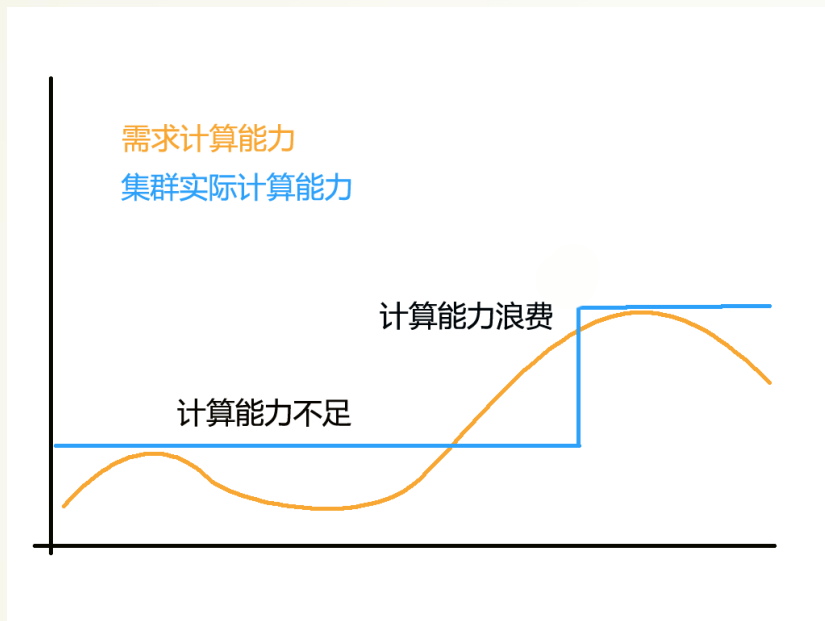
存储扩展性对比

| 本地集群 | Amazon S3 |
|----------------------------|---------------------|
| 在一定时间内不可扩展 增加硬盘 购买机器 | 无限可扩展 单个文件限制小于7T |

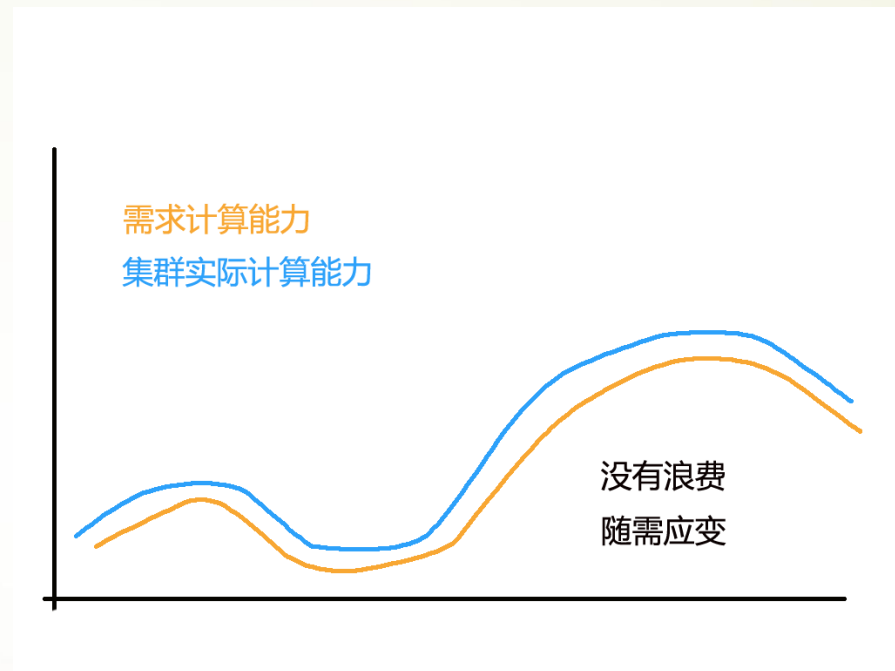
计算能力扩展性对比

| 本地集群 | Amazon EC2 |
|---------------------|--------------|
| 在一定时间内不可扩展 购买新机器 | 按需申请 自动扩展 |

按需应变的计算能力



本地 Hadoop



EMR

伸缩响应速度对比

| 本地集群 | Amazon AWS |
|-----------------------------|---------------------------|
| 慢 需走硬件采购流程 需要安装配置软件环境 | 快 可按实际需求随时扩展或者 缩小规模 |

成本对比

| 本地集群 | Amazon AWS |
|---------------------------|---|
| 一次性投入，持续折旧 资源浪费与资源不足并存 | 按需租用，消灭浪费与不足 只为使用时间付钱 费用摊分到每个使用时间 |

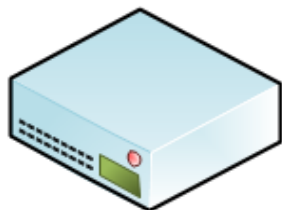
成本对比

10台服务器 = $30000 \times 10 = 300000$ 元

每月折旧成本 = $300000/5/12 = 5000$ 元

托管费用 = ?

运维人员费用 = ?



内存增强型超大实例

$\$(0.41+0.19) / \text{hour}$

内存17.1G，6.5 个EC2 计算单位，

420 GB 本地存储

END.....

Database

BDaaS

flowingdata

DB2

NoSQL

MySQL

Oracle

Big Data