SDCC 2016
中国软件开发者大会
SOFTWARE DEVELOPER CONFERENCE CHINA

Tera

百度的万亿量级数据库

颜世光
百度搜索基础架构团队

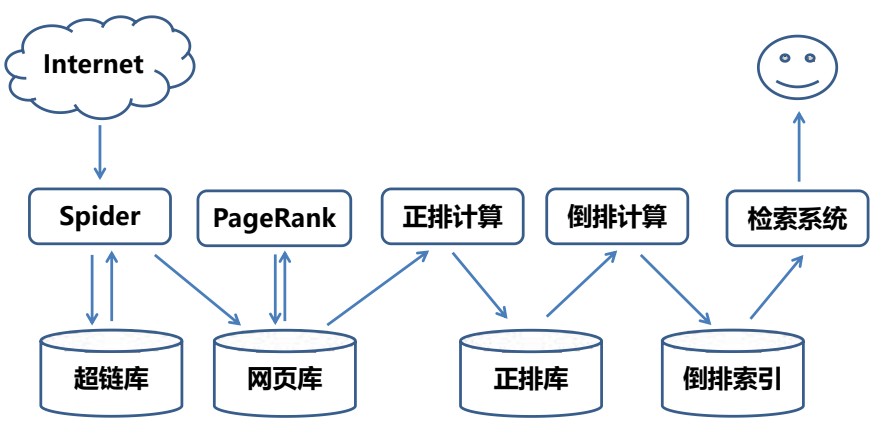
个人介绍

- 颜世光, 百度搜索基础架构Tech Lead
- 代表作品
  - 百度第三代Spider系统
  - 百度文件系统 BFS
  - 集群调度系统 Galaxy
  - 海量实时数据库 Tera
- 个人主页&Blog
  - https://github.com/bluebore
  - http://bluebore.cn

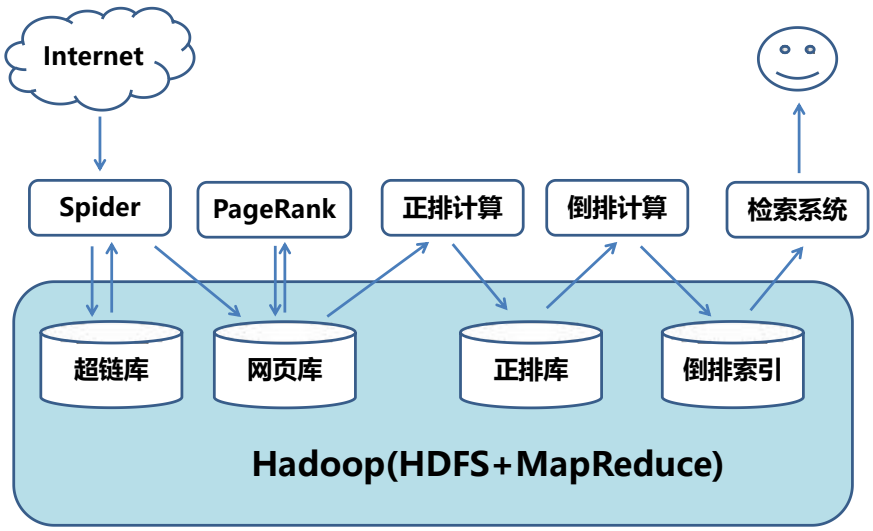
提纲

- 搜索引擎面临的挑战
- Tera的设计
- 工业实践总结
- Tera的应用
- 未来工作

搜索引擎与数据库



Tera产生之前



过去10年

- 互联网上网页数量
  - 百亿 -> 万亿
- 网民的期望
  - 新网页产生到能检索到
  - 几个周 -> 几分钟

Hadoop的功与过

- 大数据处理
  - Load balance
  - Failover
- 计算可以获得全局信息
  - 一篇网页的价值谁说了算?
  - 网页聚类、去重
  - PageRank
  - 正排转倒排

Hadoop的功与过

- 线性扩展问题
  - 50亿网页处理 -> 500台服务器
  - 5万亿网页处理 -> 50万台？
  - 解决：必须增量处理
- 时效性问题
  - 几十轮MR过程，耗时数天
  - 解决：必须流式处理

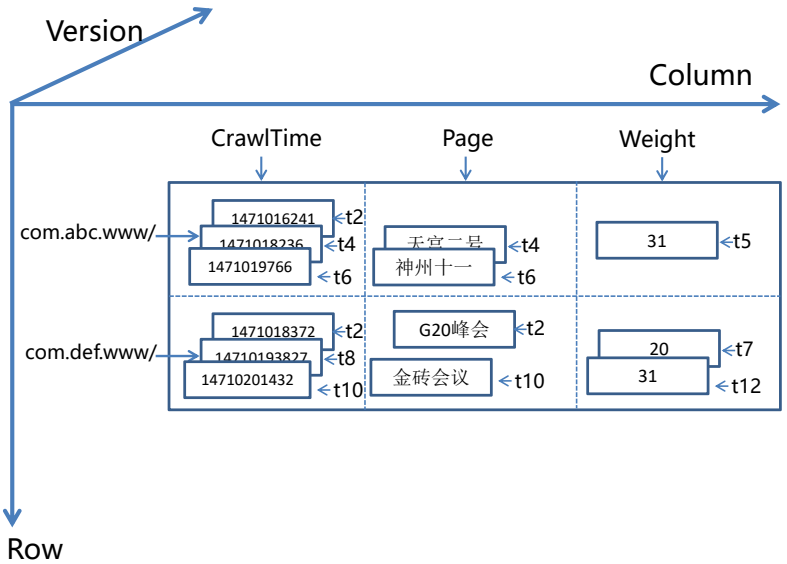
我们的解决方案

- 海量实时数据库Tera
  - 分布式、可扩展
    - 万亿记录数，百PB容量，千万QPS读写
  - 全局有序
    - 多维度统计、调度
  - 自动负载均衡
    - 互联网热点频发，业务迭代迅速
  - 多版本、表格快照
    - 历史数据分析、业务数据回滚
  - 列存储
    - 快速随机访问、分钟内扫描PB数据
  - 分布式事务

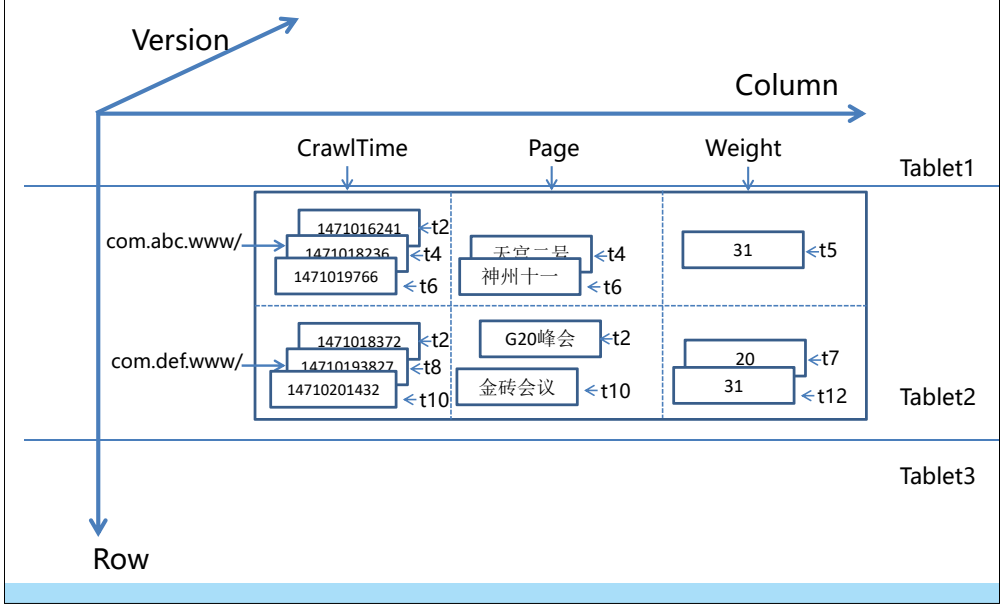
类似的解决方案

- Google
  - Bigtable
  - Percolator
  - Spanner
  - 问题: 不开源，只有理论
- Java社区
  - Apache HBase
  - Xiaomi Thems
  - 问题: 扩展性、稳定性和延迟

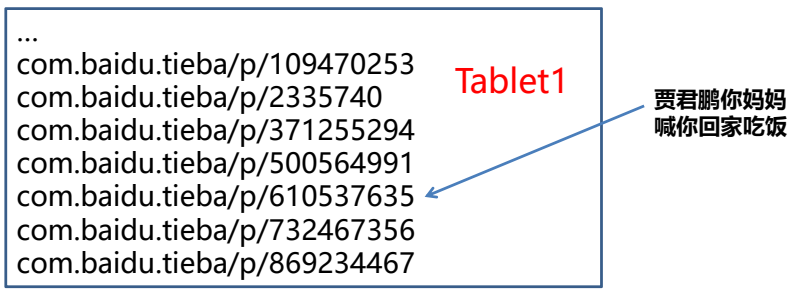
Tera数据模型



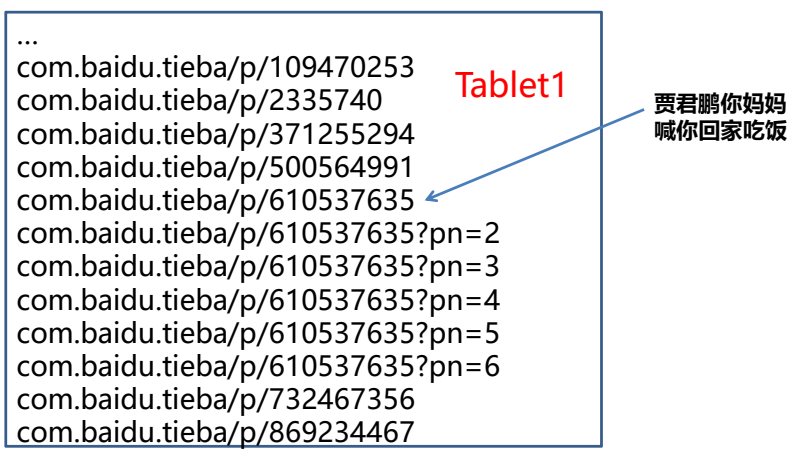
按行切分成多Tablet



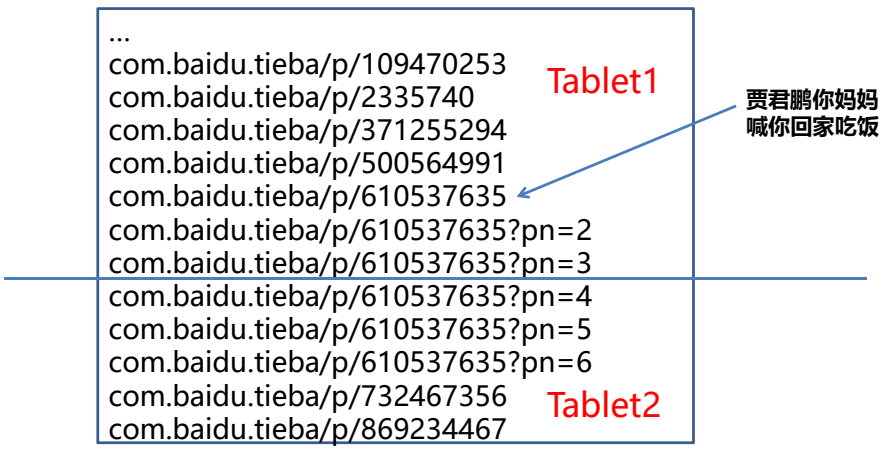
区间自动切分



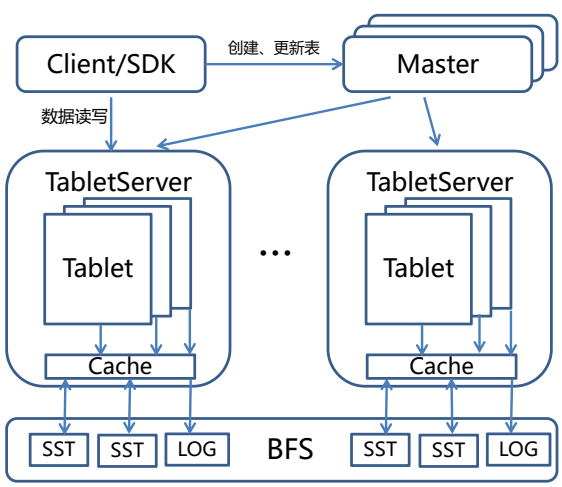
区间自动切分



区间自动切分



Tera架构

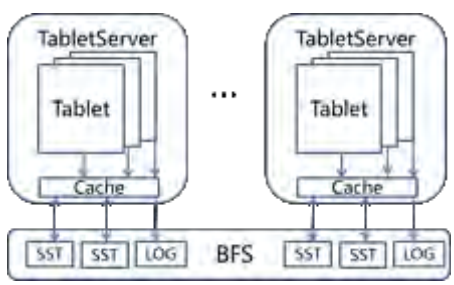


Tera给我们带来了什么?

- 数据库系统最想要的是什么?
  - 自动伸缩
    - 业务快速增长 50节点 -> 80节点
    - 促销活动 1500节点 -> 2500点
    - 活动过后 2500节点 -> 1000节点
  - 稳定可靠
    - 强一致、高可用
    - 热点自动均衡

核心技术-自动负载均衡

- 区间自动分裂
  - 快: <50ms
  - 通过文件引用实现
- 区间自动迁移
  - <50ms
  - 基于DFS，无数据拷贝
- 区间在线合并
  - 仅元数据变更，代价小，时间短(200ms)
  - 所以可以是自动的，无需人工干预

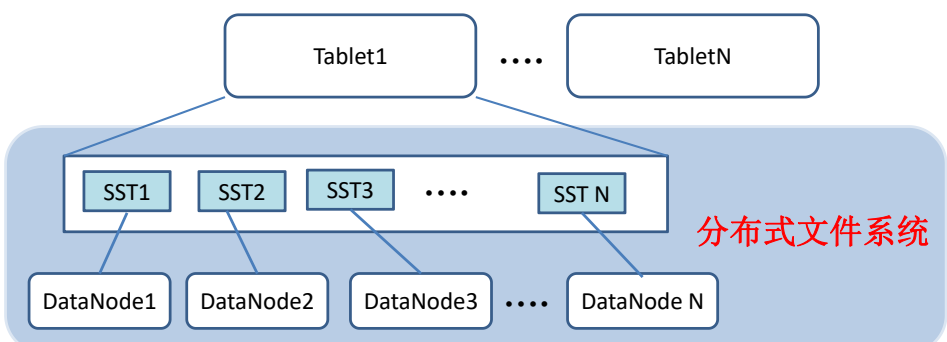


解决访问热点问题

- 传统方式 - 缓存
  - 文件块Cache
  - 数据记录Cache
  - 查询结果Cache

解决访问热点问题

- 分布式文件系统
  - 表面上：实现了快速分裂与迁移
  - 本质是：天然将请求打散到数千节点

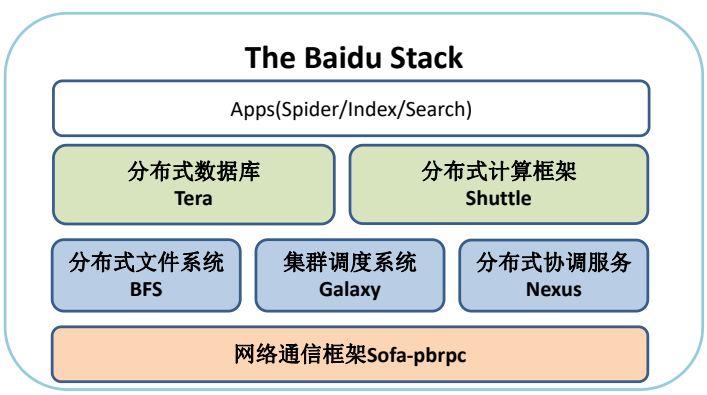


The Baidu File System

- 可用性
  - NameServer Cluster
  - 无单点
- 扩展性
  - NameNode拆分
- 性能
  - 高吞吐
    - 单机 1.1GB/S读写吞吐
  - 低延迟
    - 读写长尾优化
    - C++11实现

工业实践 – 分层设计

- 分工、复用
  - 问题最好解决一次
  - 一处解决多处受益



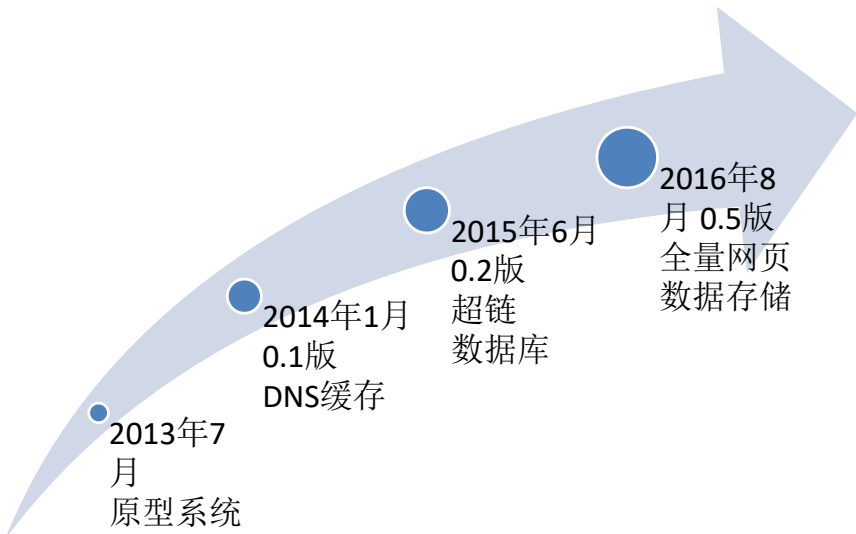
工业实践 – 可用性设计

- 硬件&软件故障不可避免
  - MTBF是30年的机器
  - 搭建一个1万台的集群
  - 每1~2天坏一台
- 降低故障恢复时间
  - 可用性 = (总时间 - 故障数 * 恢复时间) / 总时间
  - HBase 几分钟
  - Tera 几秒钟

工业实践 – 低延迟设计

- Backup Requests
  - 2ms后发送备份读请求到第二个副本
  - 如果一个被响应了，Cancel掉另外一个
  - 99.9分位延迟降低80%
- 慎用自动GC的语言
  - 实时处理, 大量小请求,频繁触发STW
  - 服务无响应
  - 不必要的failover

Tera的发展



Tera在百度的应用

场景	描述	规模
DNS缓存系统	缓存站点IP、Robots等信息	~10TB，~100亿条记录， 每天~100亿次读写访问
超链数据库	全网超链接数据 (包含数年历史信息)	~10PB，~10万亿条记录， 每条记录数百列， 每天~1万亿次读写访问
网页数据库	全网网页数据	~100PB，~1万亿条记录， 每条记录数百列 每天~1万亿次读写访问

未来工作

- 走出百度，走向社区
  - github.com/baidu/bfs
  - github.com/baidu/tera
  - 开发和Code Review都在GitHub上
  - 来自社区的贡献已经在驱动百度搜索

微信群满，交流请加我或加入
QQ群：**188471131**

谢 谢！