



中科院计算所

大数据计算技术研究与应用

查礼 (char@ict.ac.cn)

大数据管理系统部

网络数据科学与工程重点实验室

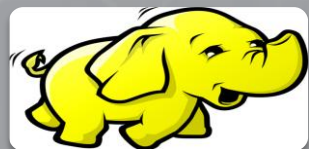
中科院计算所

2013.7.9

提纲



大数据研究背景和问题



开源软件Apache Hadoop



大数据管理系统关键技术



电子商务系统的实际应用

背景：大规模数据计算

- ❖ 通信、网络、存储、传感器等电子信息技术飞速发展导致数据规模极大增加 – Big Data
- ❖ 传统的存储并处理这些数据的技术手段遇到瓶颈

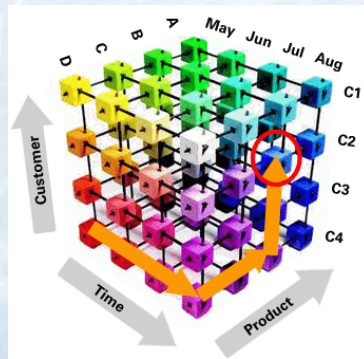
数据为王

Processing 100TB datasets

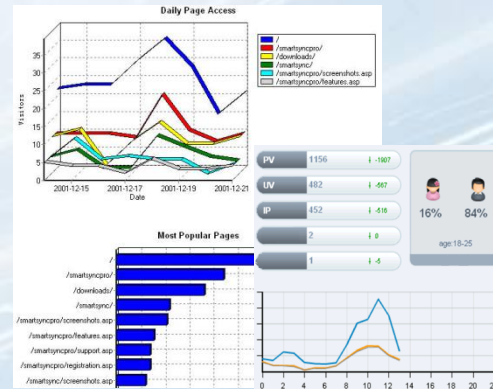
One node	Scanning@50MB/s = 35,000 min
1000 node	Scanning@50MB/s=35 min



Search Engine



Data Warehousing



Log Processing/User Behavior Analyzing



Online Data Analysis

背景 – 大数据的大问题

数据爆炸对数据存储与处理效能提出了挑战！

- Facebook用户每天上传**3亿**张照片，超过**500TB**的数据增长量，**100PB**单集群存储容量
- Google索引的在线数据2002年是5EB，到2009年增长到**280EB**
- 淘宝网注册用户达到**3.7亿**，在线商品数达到**9亿**，**14PB**海量数据存储

Growing by a Factor of 44

- 高速发展的数据型互联网企业需要**连续的系统扩展能力**

➤ 数据快速增长与数据中心扩容周期缓慢的矛盾

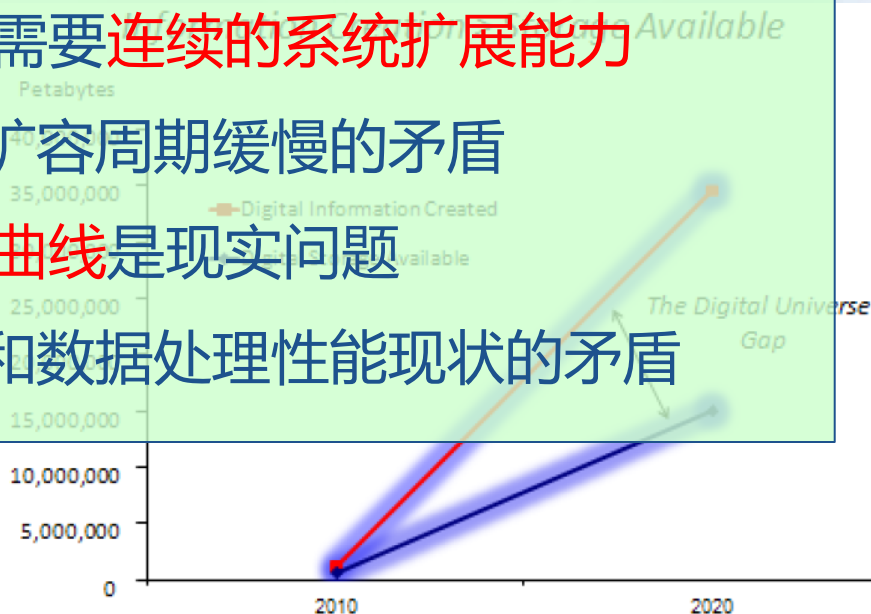
- 如何维持**低成本曲线**和**高性能曲线**是现实问题

➤ 数据业务深度的不断加强和数据处理性能现状的矛盾

2009
0.8 ZB*

* Zettabyte = 1 trillion gigabytes

2020年：
数据量将达到35ZB，较2009年增大44倍
来源：IDC Digital Universe Study, May 2010



2020年：
60%以上的创造数据将因无法存储而丢失



背景：解决大数据问题的思路

IEEE Spectrum 2009

COOLING: High-efficiency water-based cooling systems—less energy-intensive than traditional chillers—circulate cold water through the containers to remove heat, eliminating the need for air-conditioned rooms.

STRUCTURE: A 24 000-square-meter facility houses 400 containers. Delivered by trucks, the containers attach to a spine infrastructure that feeds network connectivity, power, and water. The data center has no conventional raised floors.

POWER: Two power substations feed a total of 300 megawatts to the data center, with 200 MW used for computing equipment and 100 MW for cooling and electrical losses. Batteries and generators provide backup power.

快

海量数据计算

海量数据存储

省

The Million-Server Data Center

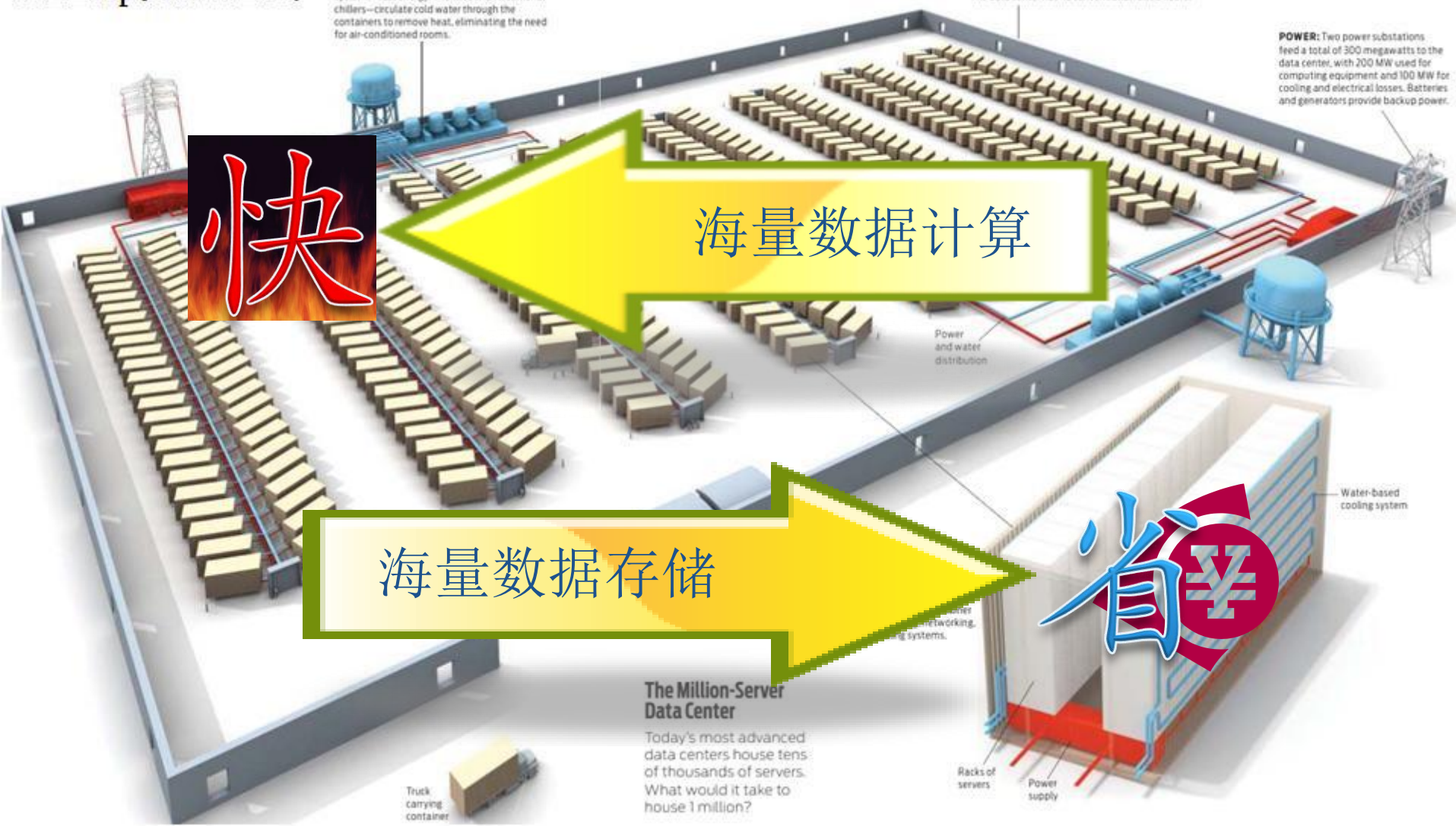
Today's most advanced data centers house tens of thousands of servers. What would it take to house 1 million?

Truck carrying container

Racks of servers

Power supply

Water-based cooling system



Facebook应用案例



Business intelligence



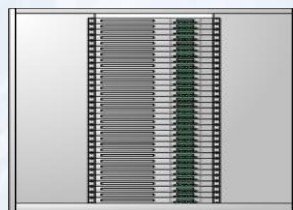
Hive Web Interface / CLI
Data Analysis/Data Mining



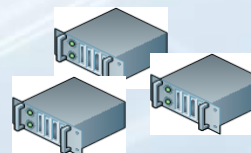
Analytics for users



Oracle RAC



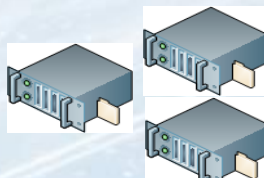
Hive/HBase on top of **Hadoop**



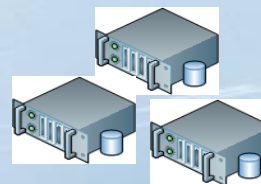
MySQL(user info)



ETL,
Workflow
Management



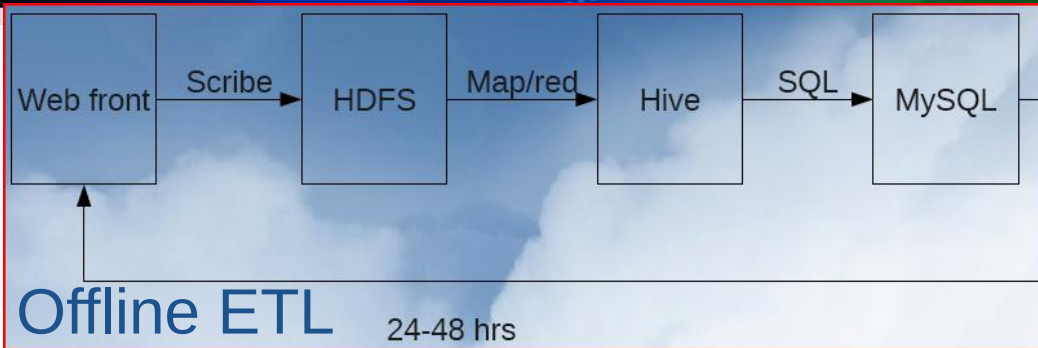
View/Click/Events Log
Storage



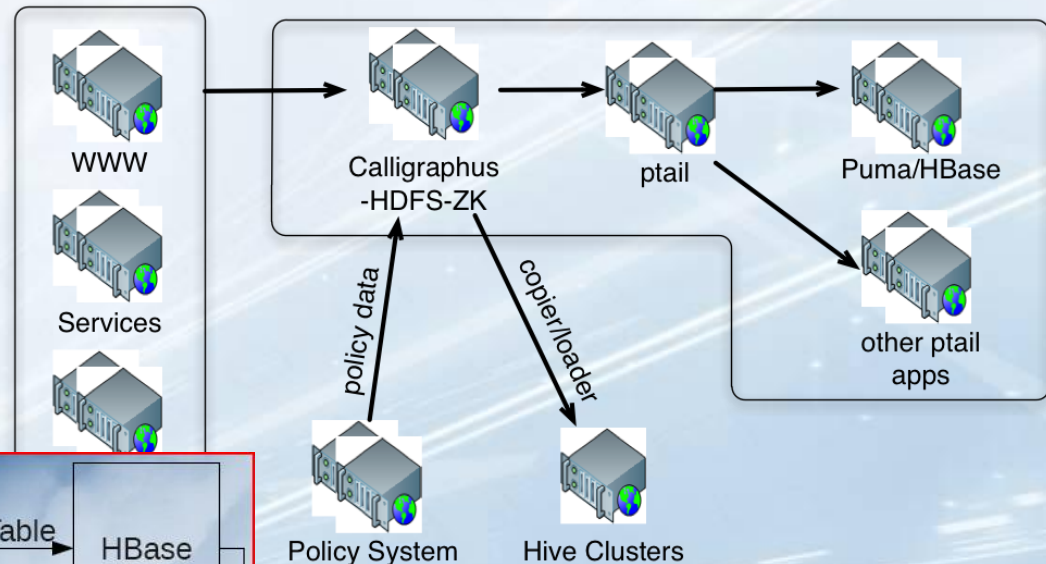
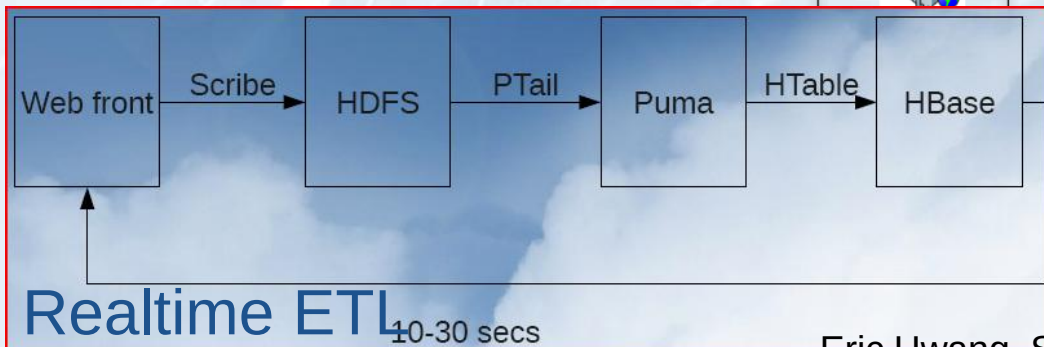
www.facebook.com



Facebook应用案例



- Scalability: 10-15 GBytes/second
- Reliability: No single point of failure
- Data loss SLA: 0.01%
 - loss due to hardware: means at most 1 out of 10,000 machines can lose data
- Delay of less than 10 sec for 99% of data
 - Typically we see 2s
- Easy to use:
 - as simple as 'tail -f /var/log/my-log-file'



Big Data Landscape

Vertical Apps



Ad/Media Apps



Business Intelligence



Analytics and Visualization



Log Data Apps



Data As A Service



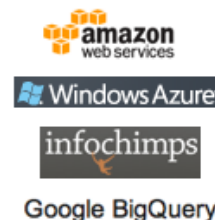
Analytics Infrastructure



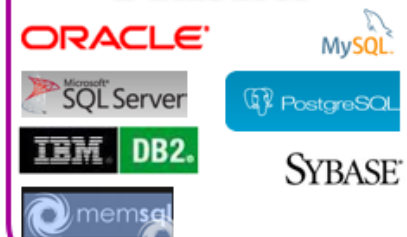
Operational Infrastructure



Infrastructure As A Service



Structured Databases



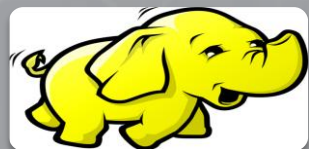
Technologies



提纲



大数据研究背景和问题



开源软件Apache Hadoop



大数据管理系统关键技术



电子商务系统的实际应用

Hadoop

- ❖ Apache Nutch, 2002
- ❖ NDFFS + MapReduce, 2004
- ❖ Hadoop, 2006
- ❖ Apache Hadoop, 2008

- <http://hadoop.apache.org/>

- Book:

- <http://oreilly.com/catalog/9780596521998/index.html>

➤ **Clone of Google's GFS and MapReduce**

- Can process large scale Web pages

➤ **Written in Java**

- Does work with other languages

➤ **Runs on**

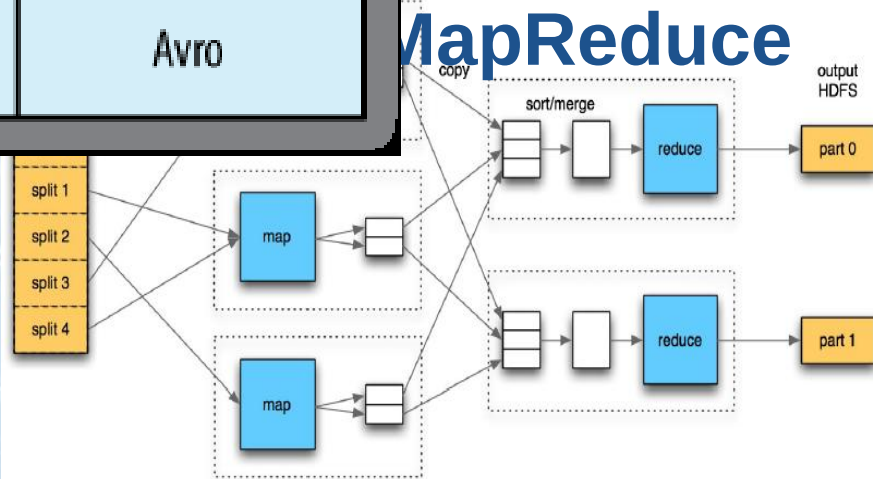
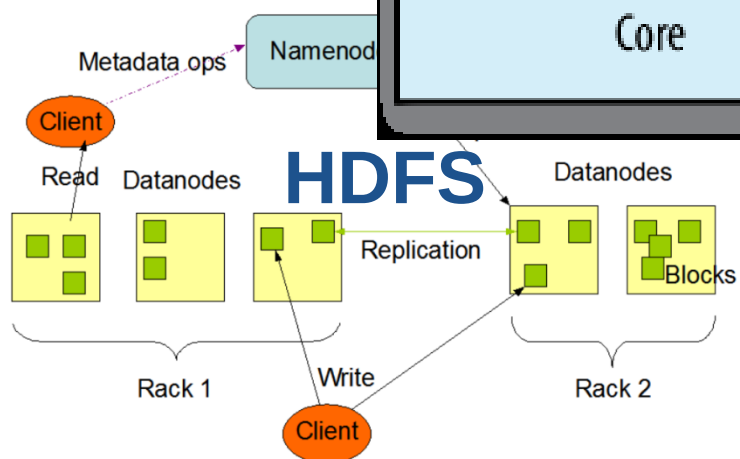
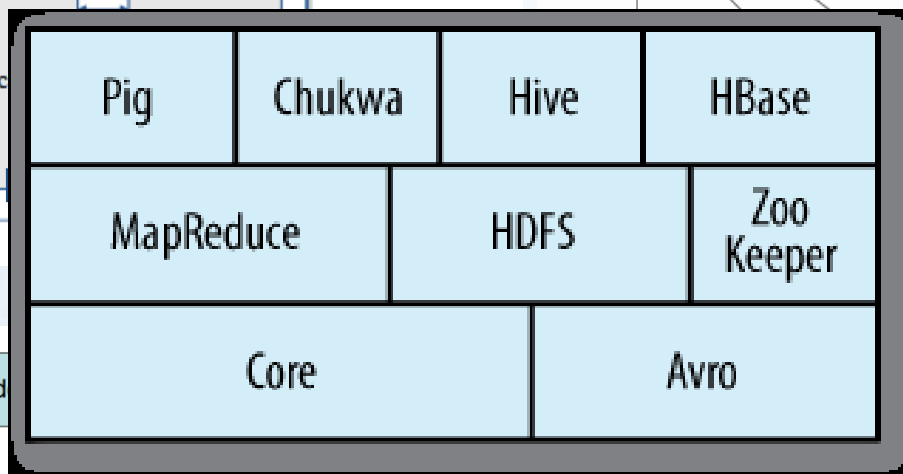
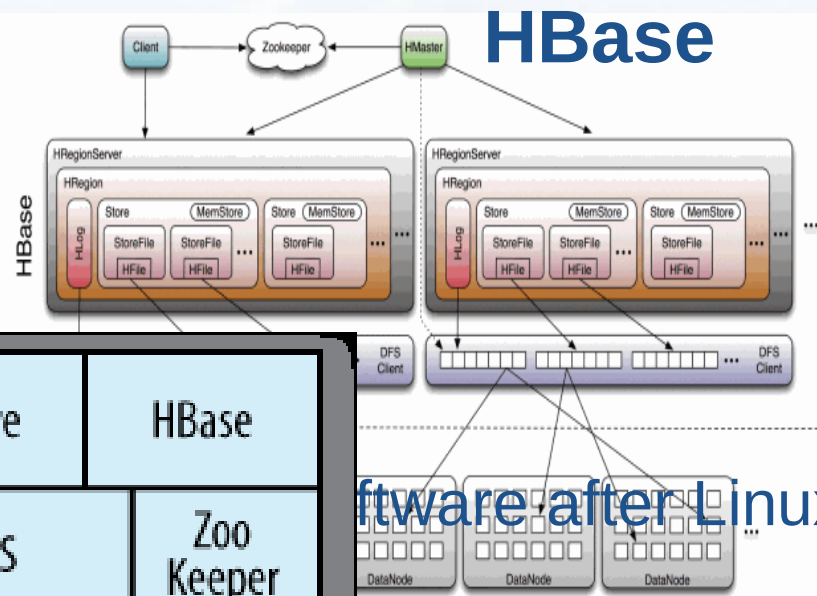
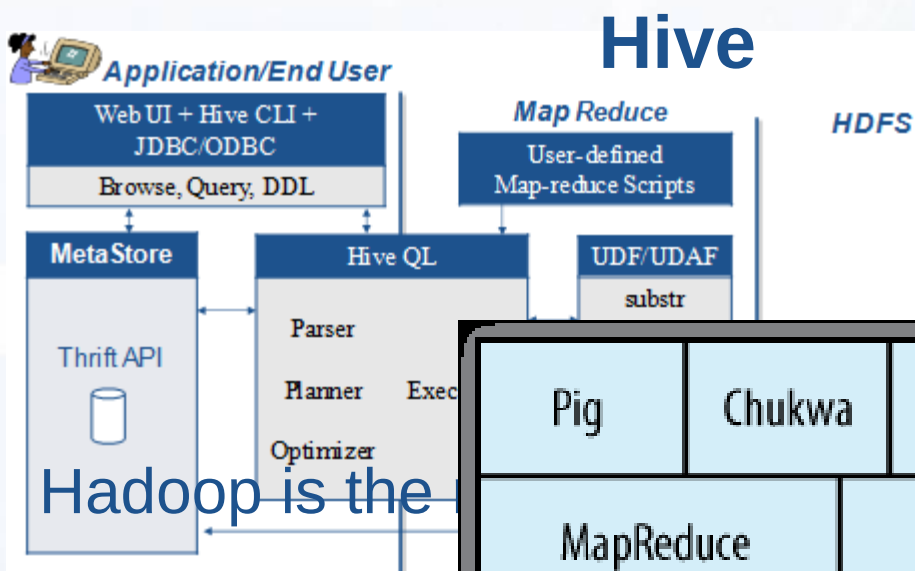
- Linux, Windows and more
- Commodity hardware with high failure rate



Doug Cutting,
Apache软件基
金会主席



Hadoop 组成部分



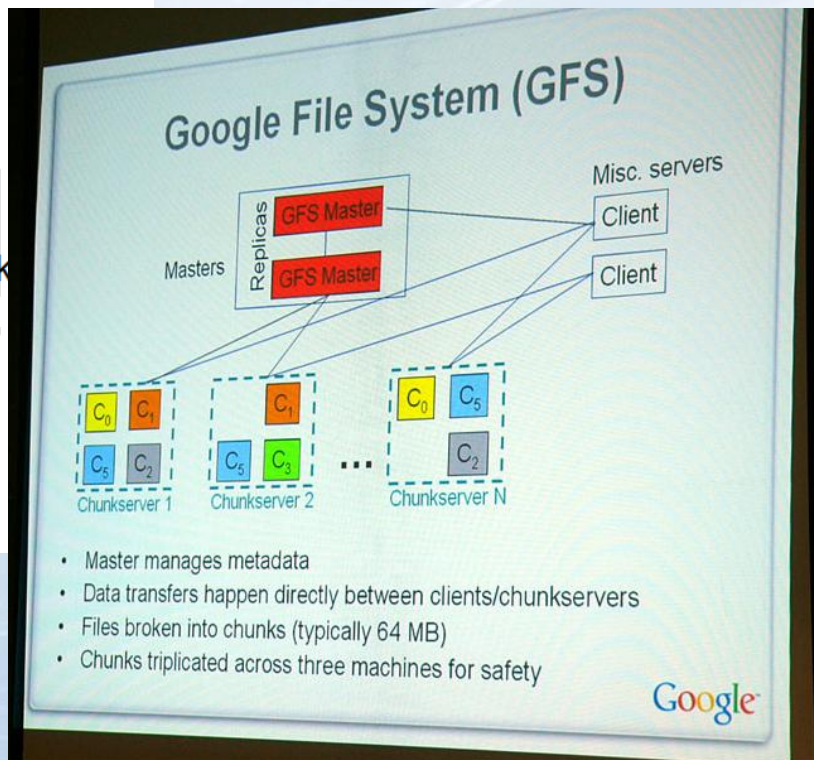
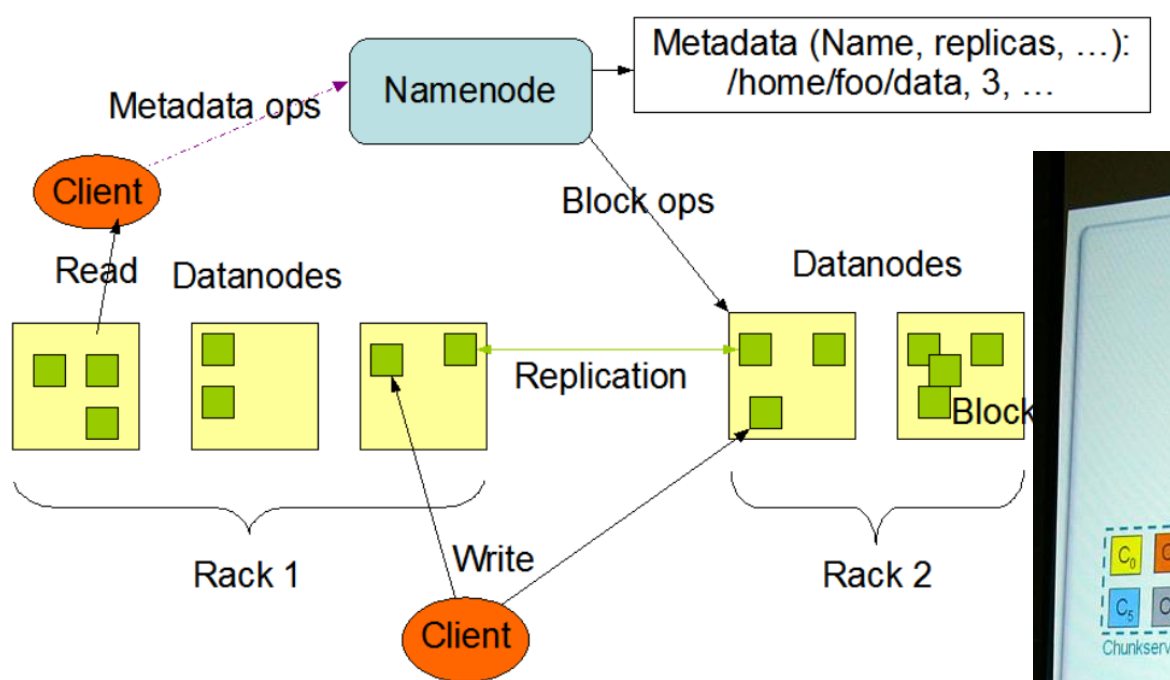
software after Linux.

Hadoop HDFS体系结构

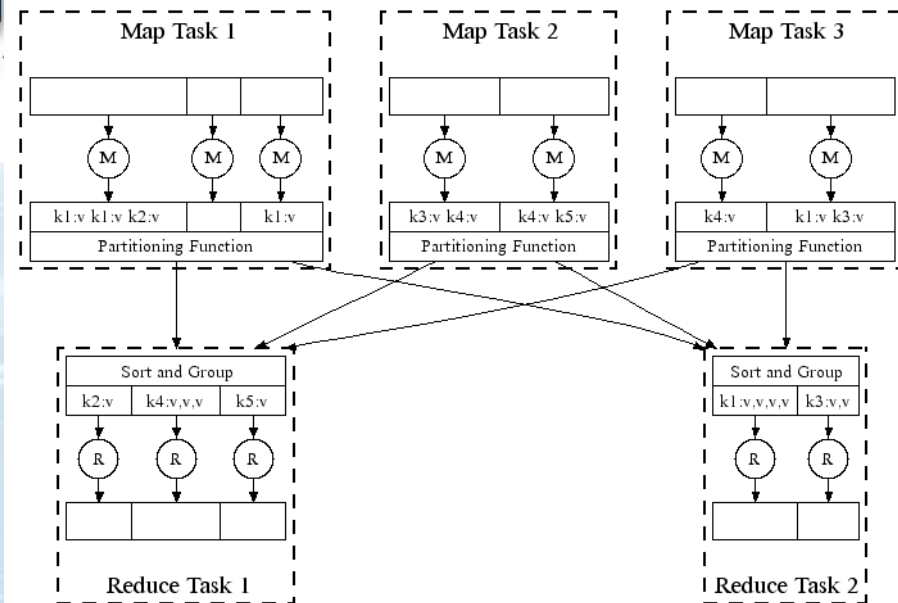
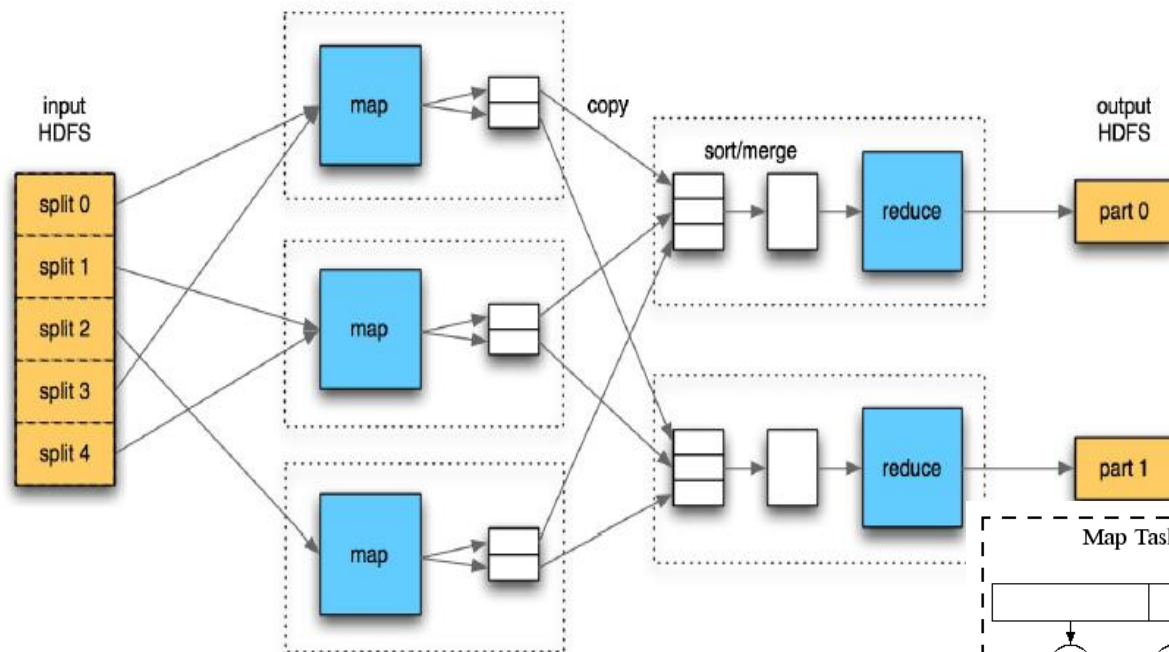
规模: 10K nodes, 100 million files, 10 PB

特性：适合数据批处理；最大化吞吐率；允许计算向数据迁移

优化：数据块副本、数据块放置策略、缓存策略等



Hadoop MapReduce处理流程

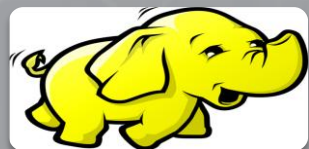


Dean & Ghemawat: "MapReduce: Simplified Data Processing on Large Clusters", OSDI 2004

提纲



大数据研究背景和问题



开源软件Apache Hadoop



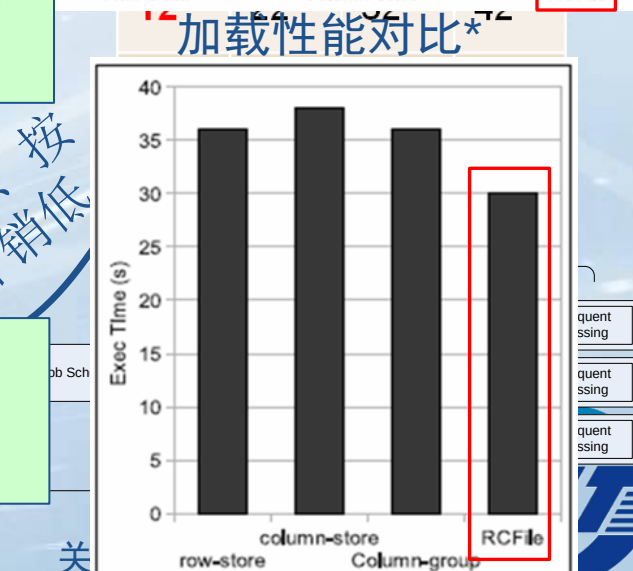
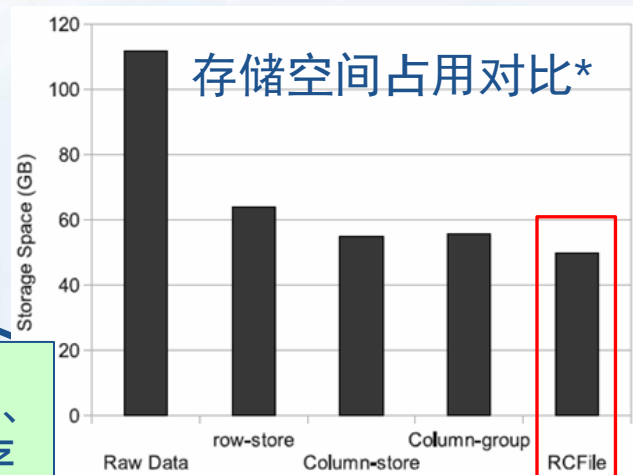
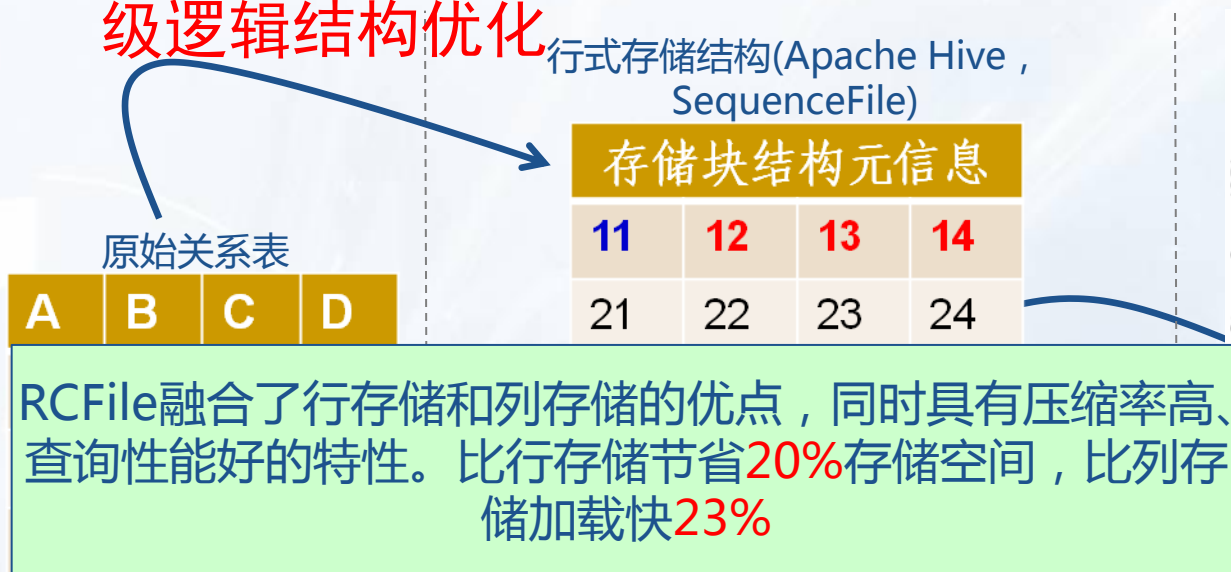
大数据管理系统关键技术



电子商务系统的实际应用

行列混合式数据存储技术RCFile

- RCFile将关系数据水平分块，块内按列序存储，实现文件级逻辑结构优化

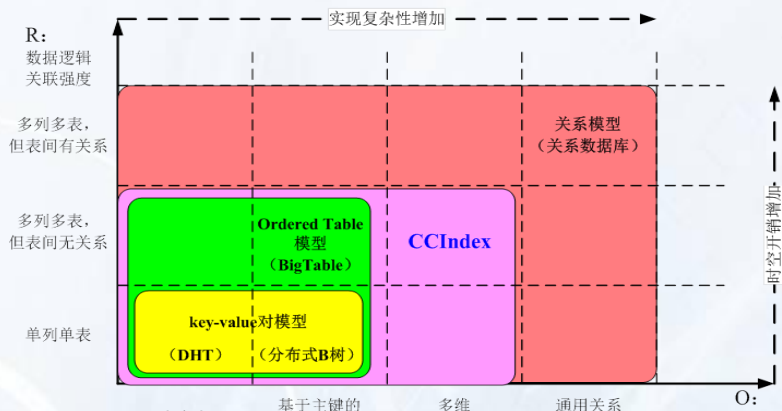


解决了海量数据的空间占用和数据加载共性问题

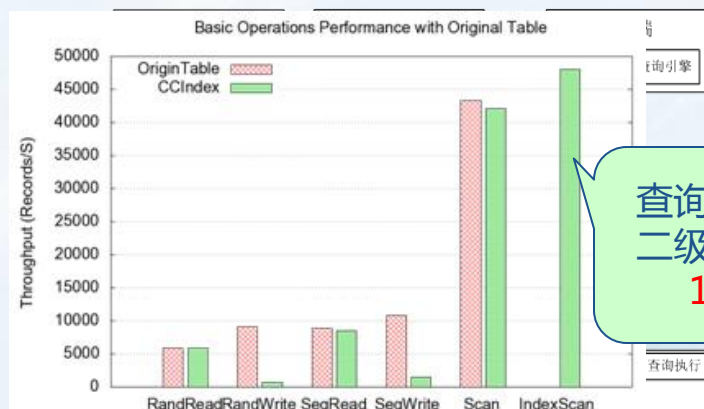
*采用Facebook真实用户访问日志数据负载评测

互补式聚簇索引技术CCIndex

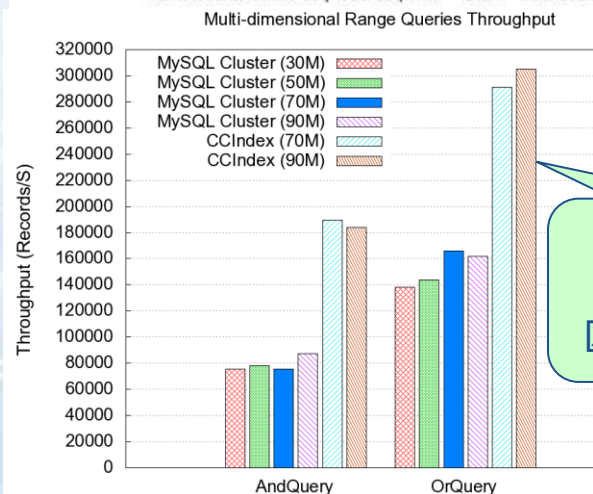
- CCIndex利用冗余的副本数据块为多数据列构建聚簇索引，不增加额外存储空间，实现数据块级布局结构优化以及分布式实时查询统计能力



CCIndex保持了BigTable数据模型高可扩展和高吞吐率特性，同时具有关系数据模型的查询统计能力



查询性能是常用二级索引技术的10倍以上



比MySQL Cluster内存数据库的多维区间查询性能快1倍

解决了海量数据实时分析计算的共性问题

互补式聚簇索引技术(直接扫描对应的聚簇索引表)

按字典序排列

A1 < A2 < A3 < A4

B4 < B3 < B2 < B1

二级索引技术(先扫描对应索引表再查原表)

CCIndex将二级索引变为一索引，不访问原表直接进行区间查询和统计

示例: select count(c1) from TAB where c1 < A and c2 > B

RCFile技术性能优势和应用情况

Certification on Adoption of RCFile Technology

17 October 2010

To Institute of Computing Technology, Chinese Academy of Sciences

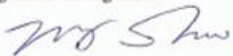
Apache Hive is a data warehouse infrastructure built on top of Apache Hadoop. It enables easy distributed big data management, and provides the capability to query and analyze large data sets stored on Apache Hadoop. Apache Hive is widely adopted and constructs the distributed big data infrastructure inside Facebook, Digg, CNET, Taobao, and other famous large-scale internet companies [Reference: <https://cwiki.apache.org/confluence/display/Hive/PoweredBy>].

The RCFile technology invented by Institute of Computing Technology, Chinese Academy of Sciences was contributed Apache Hive project with Apache License on 5/1/2009 [Reference: <https://issues.apache.org/jira/browse/HIVE-352>]. RCFile combines advantages of both row store and column store. High compression ratio on column data can reduce the usage of disk space and the horizontally split group can decrease the overhead of data loading stage. Compared with SequenceFile, which was the row store technology in Apache Hive, RCFile can achieve up to 20% space savings without affecting query performance. Compared with column group technology used in Apache Pig, which is another big data analysis system, RCFile's data loading is 23% faster as far as the disk space utilization ratio is almost equal. [Reference: <http://en.wikipedia.org/wiki/RCFile>] RCFile has been integrated into Apache Hive since version 0.4.0, and has replaced SequenceFile as the default binary data storage structure. According to our knowledge, a great number of internet companies world-wide have begun to use RCFile store their data since year 2009. Obviously, RCFile has become the de facto standard of data storage structure inside distributed offline data analysis systems such as Apache Hive.

I can fully certify the authenticity of above statements about RCFile and its integration to Apache Hive project. Please contact with me in case of any questions.

Sincerely yours,

Signature: Zheng Shao



Project Management Committee member and committer

Apache Hive Project

Email: zshao@apache.org

Phone: +1-650-353-0457

Address: 1076 Tanland Dr Apt 103, Palo Alto, CA 94303, USA

- Compared with SequenceFile, which was the default row store technology in Apache Hive, RCFile can achieve up to 20% space savings without affecting query performance.

- Compared with column group technology used in Apache Pig, which is another big data analysis system, RCFile's data loading is 23% faster as far as the disk space utilization ratio is almost equal.

- Obviously, RCFile has become the de facto standard of data storage structure inside distributed offline data analysis systems such as Apache Hive.

CCIndex技术性能优势和应用情况

用户证明

中国科学院计算技术研究所：

淘宝网作为亚洲地区最大的电子商务公司，致力于海量数据的分析挖掘和可视化展现，为电子商务用户提供专业权威的数据产品和服务。数据魔方是淘宝网首个基于全网数据的数据产品，是淘宝从电子商务公司向深度数据服务公司转型的里程碑式的产品。数据魔方通过对海量数据的分析与挖掘，能够让企业了解行业市场发展趋势、发现市场最新热点、掌握竞争对手动态、洞察买家购物习惯。目前，数据魔方在广度上覆盖了淘宝网 60 多个行业、8000 多个类目、5.9 万品牌、8.8 亿在线商品的真实交易数据。其特色功能包括：市场研究分析，店铺分析，消费者分析，营销效果分析，品牌市场占有率分析，消费者对品牌和产品的关注度分析，热门品牌与热门产品排行等。

面向海量结构化数据的在线式多列区间查询是深度数据分析与服务的重要需求，由中国科学院计算技术研究所研发的互补式聚类索引（Complemental Clustering Index, CCIndex）可以解决诸如 HBase 等当前主流的列簇式 NoSQL 数据库系统在多列查询上的功能缺失和性能低下的问题。

结合数据魔方在线系统的实际需求，作为数据魔方全属性实时计算系统的核心，CCIndex 技术经适配和优化后，已在我公司上线投入实际运行。CCIndex 使我公司原有的业务逻辑能够直接迁移到经改进的 HBase 上，同时增强了全属性实时计算系统的扩展性和性能。目前，实时计算系统处理的数据条目超过 108 亿。采用 CCIndex 技术后，系统处理的数据时效范围从原来的 7 天增大到 3 个月，处理的数据规模增大了 7 倍，系统吞吐率增大了 7 倍，对原来延迟大于 1s 的查询请求平均降低了 57.4%。



CCIndex技术可以解决诸如HBase等当前主流的列簇式NoSQL数据库在**多列查询上的功能缺失和性能低下**的问题。

采用CCIndex后，在硬件规模保持不变的前提下，系统处理的数据时效范围从原来的7天增大到**3个月**，处理的数据规模增大了**一个数量级**，系统吞吐率增大了**7倍**，对原来延迟大于1s的查询请求响应时间平均降低了**57.4%**。

与国外同类技术比较

技术指标	本项目	国际领先系统
数据存储空间占用	RCFile比行存储 节省约20%空间	Apache Hive (SequenceFile)
数据加载性能	RCFile比列存储 加载快23%以上	Apache Pig (Zebra)
在线查询分析性能	CCIndex是二级 索引技术的11.4倍	Apache HBase (IndexTable)

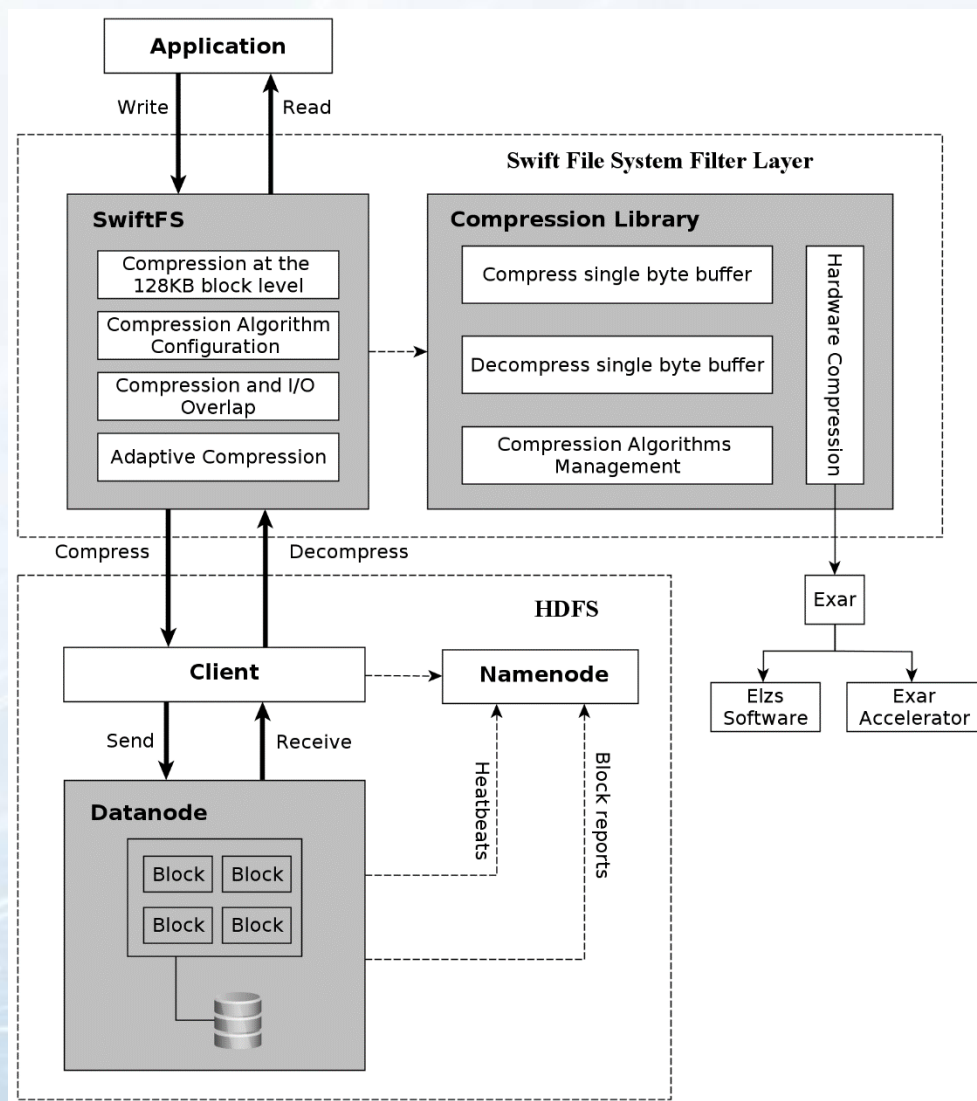
实现PB级数据离线分析处理和百亿记录级数据实时查询分析，成为国际事实标准



用户透明的HDFS数据压缩解压缩

- ❖ 自动将文件分割成多个固定大小的数据块，每个数据块单独压缩
- ❖ 每一个文件由一个压缩文件和索引文件构成
- ❖ 可应用于Hadoop平台上的文件系统

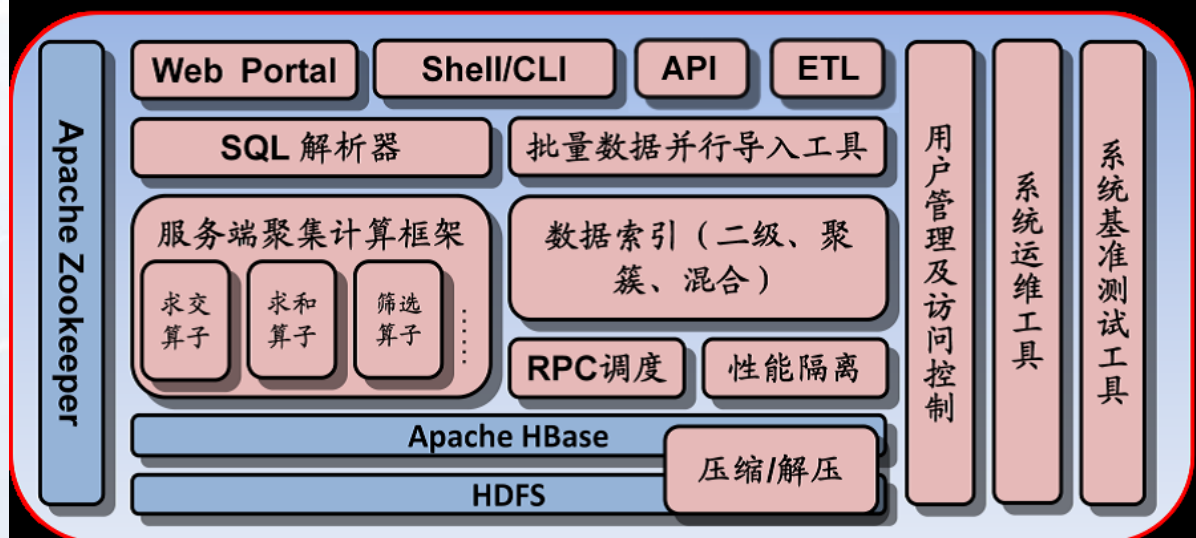
在不影响其它文件系统的情况下，以动态、透明的方式给文件系统添加压缩功能



技术特色

- ❖ 极高的性能，高通量数据存取，分布实时分析计算，TB级数据支持，百倍优于传统关系型数据库。
- ❖ 软硬一体。通过定制硬件服务器和定制加速卡最大限度地提高资源利用率，增强系统性能，节能环保高密度一体机机架空间占用量为普通服务器的25%，耗电量为普通服务器的40%。
- ❖ 数据接入简单，提供JDBC、ETL等通用数据接口，应用对接远优于新兴的NoSQL数据库，多种管理套件应用维护简单。

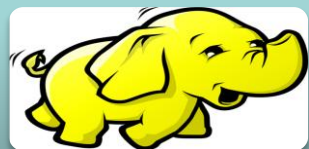
多租户、高并发、低延迟在线式数据查询分析负载



提纲



大数据研究背景和问题



开源软件Apache Hadoop



大数据管理系统关键技术



电子商务系统的实际应用

淘宝数据魔方应用

查询条件

所有分

行业名称: 女装/女士

使用进度:

还剩8天到期

市场分析

市场整体行情

- 热销品牌排行
- 热销产品排行
- 热销宝贝排行 **HOT**
- 热销特征排行
- 关键词成交分析 **HOT**

买家分析

- 买家规模分析
- 买家地域分析
- 购买时段分析
- 来访时段分析
- 买家信用分析
- 性别年龄分析
- 购买频次分析
- 客单价分析
- 商品标价分布 **NEW**

卖家分析

- 热卖店铺排行 **HOT**
- 卖家规模分析

HBase: Data Storage, Data Aggregation

Region Server:
Data Storage
Index+Data
Data Aggregation

Region Server:
Data Storage
Index+Data
Data Aggregation

Region Server:
Data Storage
Index+Data
Data Aggregation

.....

Query

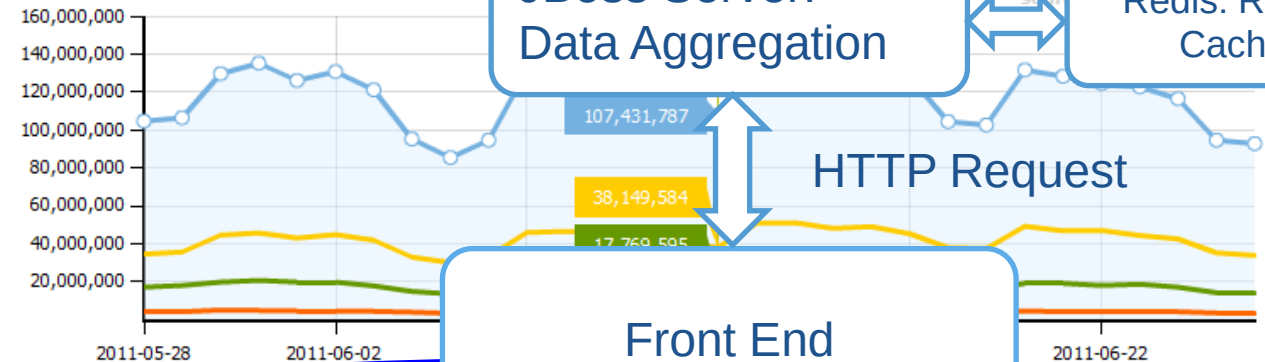
Aggregation
Result

JBoss Server:
Data Aggregation

Redis: Result
Cache

HTTP Request

Front End



☒ 女装/女士精品 : 107,431,787

☒ 连衣裙 : 3,874,984

输入不同行业对比:

☐ 连衣裙

☐ T恤

☐ 牛仔裤

开始对比

取消

淘宝数据魔方 - 全属性实时计算

数据装载

全网交易数据



云梯

HBase

HBase

属性对

索引：交易id列表

交易1 (二进制, 定长)

交易2

数据查询

求SUM(alipay)

属性	属性值
笔记本尺寸	13寸
笔记本定位	商务定位

查索引

节点1

1, 2, 4, 9

节点2

1, 4, 7

求交集

节点2

1, 4

本地SUM运算 (HBase扩展)

汇总计算
写入缓存



淘宝量子统计访客直播间-业务背景

热区分段 秒, 刷新间隔 秒, 店铺ID

日总值(PV) **1305055** (UV) **133137** 秒峰值(PV) **7** 当前值(PV) **2** 实时状态 **Running**



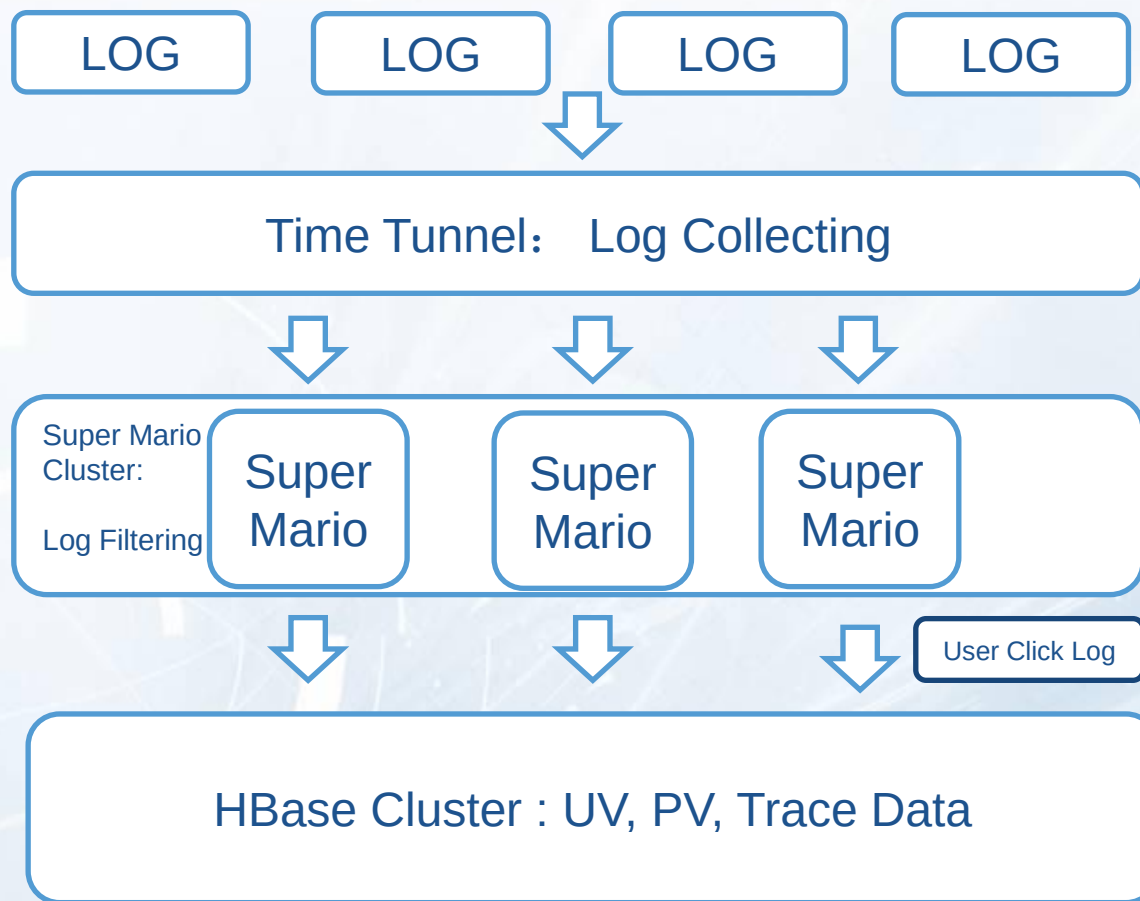
ID	Time	Source	Page	Pos	Hot	Custom
1	12:01:43	from	搜索列表页-真维斯官方旗舰店- 淘宝商城	福建省泉州市	■■■	路径分析
2	12:01:43	from	冬装 真维斯 正品 女装9针全棉圆领净色长袖棉毛衣 691505-淘宝商	河南省周口市	■■■	路径分析
3	12:01:43	from	2011新品 秋装 真维斯 正品 男装POLO领长袖T恤171206-淘宝商城	广东省佛山市	■■■■	路径分析
4	12:01:42	from	精选外套 T恤-真维斯官方旗舰店- 淘宝商城	吉林省长春市	■■■■	路径分析
5	12:01:42	from	搜索列表页-真维斯官方旗舰店- 淘宝商城	安徽省巢湖市	■■■■	路径分析
6	12:01:42	from	搜索列表页-真维斯官方旗舰店- 淘宝商城	湖南省长沙市	■■■■	路径分析
7	12:01:41	from	搜索列表页-真维斯官方旗舰店- 淘宝商城	上海市长宁区	■■■■■	路径分析
8	12:01:41	from	搜索列表页-真维斯官方旗舰店- 淘宝商城	北京市-	■■■	路径分析
9	12:01:41	from	2011新品 秋装 真维斯 正品 女装圆领中袖T恤272502-淘宝商城	湖北省黄冈市	■■■■	路径分析
10	12:01:41	from	搜索列表页-真维斯官方旗舰店- 淘宝商城	山东省烟台市	■■■	路径分析
11	12:01:41	from	搜索列表页-真维斯官方旗舰店- 淘宝商城	河北省廊坊市	■■■■	路径分析
12	12:01:40	from	2011夏装新品 真维斯 正品 女装全棉连帽横间短袖T恤673518-淘宝	陕西省西安市	■■■■■	路径分析

访客直播间

- 面向B2C 商户
- 实时用户访问情况
- 掌控用户行为, 精准营销



淘宝量子统计访客直播间-架构与效果



- 数据量级：每天12亿条，600GB
- 从用户点击到店铺展现时间延时2到3秒

数据查询及展示

The screenshot shows a web interface for data query and display. It includes filters for '地区分段' (Region Segment), '秒' (Seconds), and '店铺ID' (Store ID). The main table displays a list of data records with columns: ID, Time, Source, Page, Pos, Hot, and Custom. The table is sorted by 'Time' in descending order.

ID	Time	Source	Page	Pos	Hot	Custom
1	12:01:43	from	搜索列表页-男装和官方旗舰店-淘宝网	福建省泉州市	★	查看详情
2	12:01:43	from	女装-真维斯 正品 女装设计金线绣花修身长袖棉质秋 691505-淘宝网	河南省郑州市	★	查看详情
3	12:01:43	from	2011新品 秋装 真维斯 正品 男装POLO领长袖T恤171206-淘宝网	广东省佛山市	★	查看详情
4	12:01:42	from	服装外贸T恤-真维斯官方旗舰店-淘宝网	吉林省长春市	★	查看详情
5	12:01:42	from	搜索列表页-真维斯官方旗舰店-淘宝网	安徽省芜湖市	★	查看详情
6	12:01:42	from	搜索列表页-真维斯官方旗舰店-淘宝网	湖南省长沙市	★	查看详情
7	12:01:41	from	搜索列表页-真维斯官方旗舰店-淘宝网	上海市市长宁区	★	查看详情
8	12:01:41	from	搜索列表页-真维斯官方旗舰店-淘宝网	北京市-	★	查看详情
9	12:01:41	from	2011新品 秋装 真维斯 正品 女装圆领中袖T恤272502-淘宝网	湖北省黄冈市	★	查看详情
10	12:01:41	from	搜索列表页-真维斯官方旗舰店-淘宝网	山东省烟台市	★	查看详情
11	12:01:41	from	搜索列表页-真维斯官方旗舰店-淘宝网	河北省廊坊市	★	查看详情
12	12:01:40	from	2011夏装新品 真维斯 正品 女装全棉圆领短袖T恤875518-淘宝网	陕西省西安市	★	查看详情

腾讯广点通-业务背景-挑战

广点通：

腾讯公司为第三方应用开发商提供广告投放系统（广点通），用户通过系统可以在空间及朋友平台的多个广告位上进行应用以及应用活动相关的精准推广；此广告投放系统日收入约为200万RMB左右。

- 一秒30000次请求（30%流量，峰值）
- 一次请求200个候选广告数据
- 一个候选广告涉及7种CTR数据
- 一种CTR数据800字节：48小时的点击曝光计数
- 一次请求响应时间小于40ms
- 一次请求响应存储查询量次数 $7 \times 200 = 1400$ 次
- 一次请求响应存储查询数据量 $7 \times 200 \times 800 = 1.06\text{MB}$
- 一秒存储层查询次数 $30000 \times 200 \times 7 = 42000000$ 次/秒
- 一秒数据请求量 $30000 \times 200 \times 7 \times 800 = 31\text{GB/秒}$ （约400张千兆卡）

备注：

1、现网实时查询系统，2012.8.16日总流量12.1亿，峰值2.4w/s，均值1.4w/s，为1.7倍

2、测试设备：350C1 + 50 B6



腾讯广点通-实验系统架构

业务侧广告投放系统

HBase Client

Simple RPC

HINA

点击/曝光日志实时采集
用户行为实时采集

流式计算
清洗日志
(10台)

HBase
RPC

Ad
Recommendati
on Request

HBase Cluster

RS

Simple
RPC

In-Memory
Cache

Cache
Manager

Region for Pos

Log Table

Pos:n%20=0

RS

Simple
RPC

In-Memory
Cache

Cache
Manager

Region for Pos

Log Table

Pos:n%20=1

RS

Simple
RPC

In-Memory
Cache

Cache
Manager

Region for Pos

Log Table

Pos:n%20=2

RS

Simple
RPC

In-Memory
Cache

Cache
Manager

Region for Pos

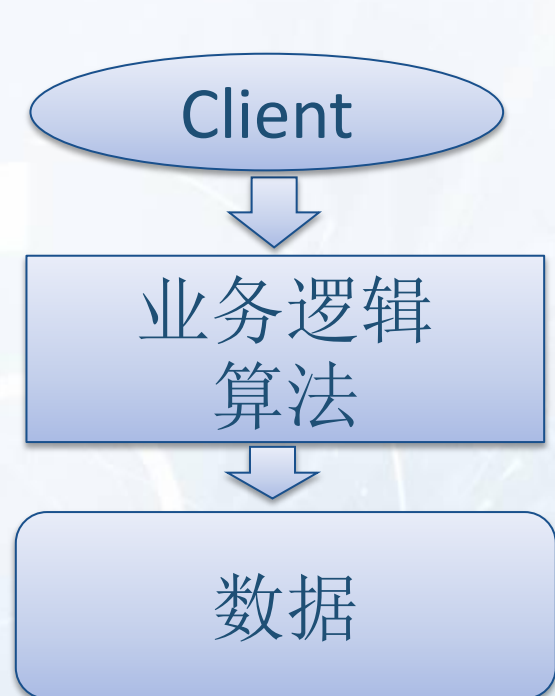
Log Table

Pos:n%20=19

数据冗余，压缩层次



广点通方案效果对比



160B6, 支撑10亿流量

原方案VS天玑Base方案



10B6, 支撑30亿流量

应用效果：

- 单服务器效率增强48倍
- 新系统日均处理日志数量30亿条，处理用户请求量25亿次，并有能力处理复杂用户推荐算法-LR算法



部分天玑Base的用户证明

用户证明



中国科学院计算技术研究所：

淘宝网作为亚洲地区最大的电子商务公司，致力于海量数据的分析挖掘和可视化展现，为电子商务用户提供专业权威的数据产品和服务。数据魔方是淘宝网首个基于全网数据的数据产品。

品，是淘
数据的分
态、洞察
5.9万品
费者分析
牌与热门
面向
学院计算
可以解决
低下的问
结合
技术经适
直接迁移
时计算系
下，系统处
系统吞吐

用户报告

中国科学院计算技术研究所：

中国科学院计算技术研究所大规模数据处理的系统结构研究方向上处于国内领先地位，其研制的流水线式资源感知任务调度算法、元信息服务器高可用技术、以及面向用户分组的层次式资源管理技术能够解决目前华为公司从TB到PB级数据处理系统的相关问题。

通过技术合作项目（YBSP2010014），相关技术应用在华为公司 iSAP 平台产品的 PB 级数据处理系统中，达到了项目的预期指标（采用流水线式资源感知任务调度算法比公平调度算法系统吞吐量提高 15.57%。采用元信息服务器高可用技术比单元信息服务器在读写混合负载下吞吐量达到原来的 132%，纯读负载下吞吐量达到原来的 179%。同时，实现了资源按部门分组隔离及组内共享机制。）

（注：本报告内容为保密信息，仅作为中国科学院计算技术研究所科研成果鉴定之用，未经允许请勿对外披露。）



地址：北京

中华人民共和国公安部

用户证明

中国科学院计算技术研究所：

海量数据存储与分析是互联网信息处理和数据挖掘的重要技术基础之一。中国科学院计算技术研究所海量数据计算领域处理大规模分布式数据处理系统业务起到重要作用，特此

中华人民共和国公安部

感谢信

中科院计算所：

2012年8月14日6时50分，苏湘渝系列持枪抢劫杀人案案犯、公安部A级通缉犯周克华被重庆公安民警成功击毙，此系列案件告破。案件的及时告破，极大地稳定了人心，鼓舞了士气，消除了这一重大社会安全隐患。

在苏湘渝系列持枪抢劫杀人案侦破过程中，贵所在技术上向我局提供了极大地帮助。在网络数据科学与工程研究中心程学旗主任的亲自带领下，查礼、刘佳、沈华伟等同志克服任务紧、难度大等困难，利用“天玑”大规模分布式数据处理系统，筛查海量网络数据，勾画犯罪嫌疑人活动轨迹，对侦查犯罪嫌疑人逃跑方向发挥了重要作用。在此，特向贵所表示衷心感谢并致以崇高敬意！希望贵所对我局工作继续大力支持！



总结

- ❖ 数据计算技术的（规模）大、（计算）快、（检索）准的发展态势
- ❖ 在计算向数据迁移的计算模式前提下，有很多问题值得重新研究，将催生如Hadoop的新技术和新的市场
- ❖ 数据规模对传统RDBMS技术提出了挑战；NoSQL技术是对RDBMS技术的补充，是实际需求导致的结果
- ❖ 天玑大数据管理系统是需求导向的，是开源软件和科研创新的无缝结合，是经由实际应用考验的成熟产品



Hadoop中国大数据技术大会



2008~2012, 近两年参会人数超1千人。已成为大数据领域的国内最重要技术会议，推动了国内Hadoop开源社区发展及生态环境建设



中科院计算所

Thank You !

