



搜狗案例分析与数据分析平台介绍

培训目标



1

搜狗案例剖析

2

搜索数据介绍

3

技术选型和架构

4

WEB Server和日志采集系统



search
Sogou 搜狗

轻松搜索 成功营销

搜狗搜索背景

搜 狐
S O H U . c o m

最大中文互联网阵地



搜狗在搜狐的肩膀上成长



搜狗搜索

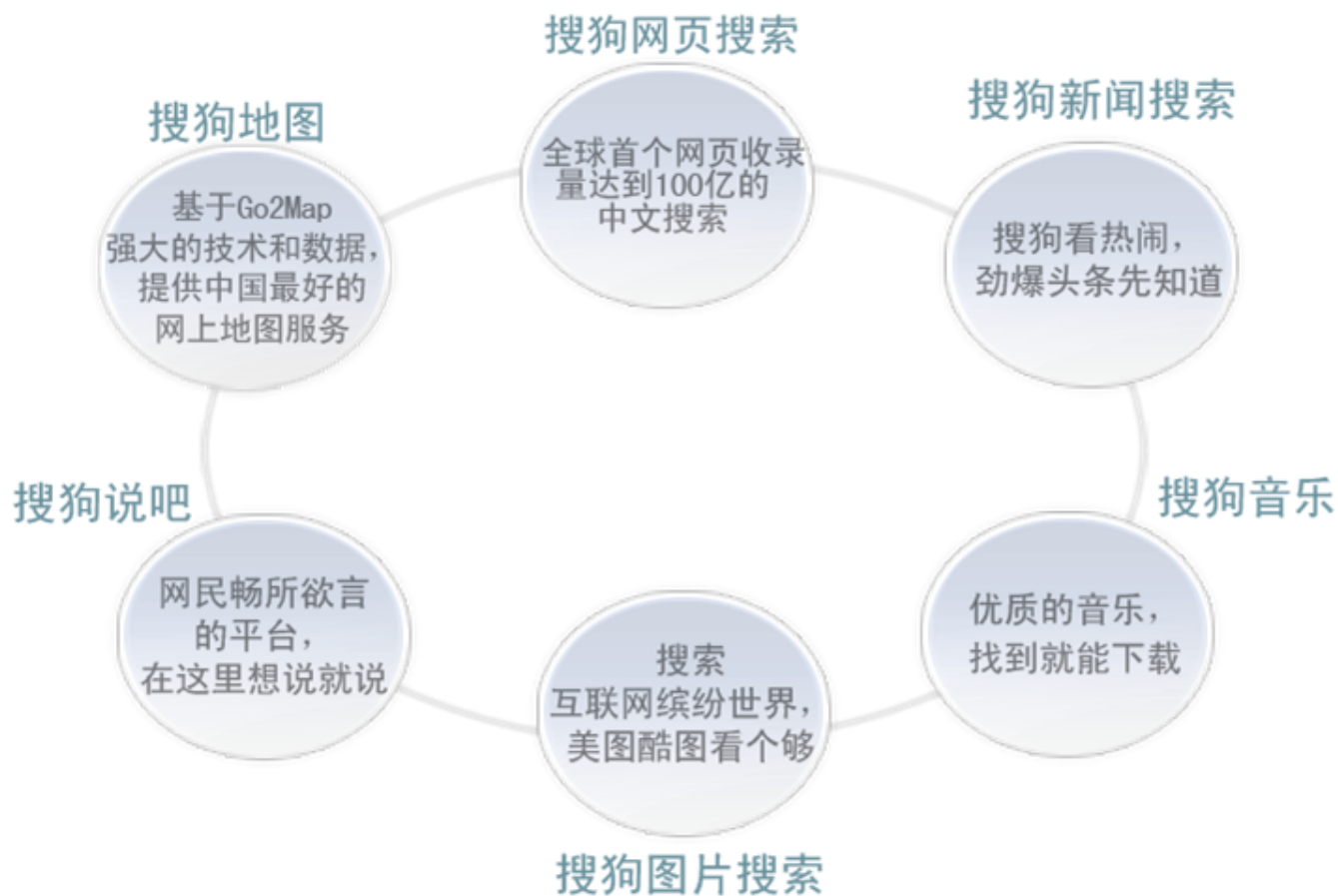
网页量200亿，日更新8亿
访问量达1.8亿

第三大中文搜索引擎

艾瑞:Q1搜索总访问次数缓增

搜狗增速显著，环比增长22%

搜狗全方位满足搜索需求



搜狗搜索 创新不断



搜狗拼音输入法

搜狗手机输入法

搜狗五笔输入法

全国网民80%以上的安装量



搜狗问答页面



新闻 网页 音乐 图片 视频 说吧 问答 更多

帮助

搜索答案

我要提问

2009 高考

在线估分 招生办主任在线答疑

大小年预测 就业查询 高校排行

搜狗问答 >> 品味生活 >> 购物

已解决

09年流行服饰大概是什么样子？

悬赏分：0 - 解决时间：2009年03月17日 21时28分

提问者：miki_005 - 超级魔法师 九级

最佳答案

格子是一直流行的款式了，比较洋气

2009年的流行的是复古装！带点欧美风！性感又带点可爱，可爱中又不失性感！比2008的流行风又上了一个层次！高贵！

答案补充

A短夹克时尚个性，叠穿也很时髦。纯粹的黑白配会显得过于扎眼，需要灰色的中和，这样不至于头重脚轻。黑色锥形裤会显得腿短，不过瘦腿到时很不错。B深色系的装束很容易显得闷，需要亮色加以调和。白色鞋子让本就矮小的身材显得更加短小。宽松马裤以及独特西装裁剪显得个性十足。C多元素的混搭最忌讳杂乱，只要有少许的重复点就足够了。这款搭配之所以比较成功，就在与色调的统一性以及元素的适当呼应。D全身的黑色装扮略显闷，但是却有拉长身形以及瘦身的功效，围巾以及帽子的运用很不错，让俊俏的脸型更显突出，简洁中有看头，是比较成功的装扮。E普通

[流行外套哪里卖的最便宜？](#)

聪明点比较购物，成就网购达人，快来smarter.com.cn
www.smarter.com.cn

[齐秦代言 野狼牛仔火爆全..](#)

狼尊国际企业以狼的精神为企业文化传票，以狼文化为品牌图腾。
www.yelang88.com

[找运动服饰项目？就上8917..](#)

好商机 新项目 赚钱项目 精选推荐 品牌服装店，更多成功选择！马上..
www.89178.com

[鱼尾纹紧致提拉弹力全能..](#)

鱼尾纹 紧致提拉 弹力全能赋活眼霜 眼袋 黑眼圈 双皇冠推荐
item.taobao.com

等待您来回答

[日本乐乐海港购物城是购物的](#)

搜狗说吧页面



[新闻](#) [网页](#) [音乐](#) [图片](#) [视频](#) [说吧](#) [问答](#) [更多](#)

穿衣搭配

搜索

☒ 进入说吧 ☐ 帖子搜索 ☐ 作者搜索

[说吧](#) >> [时尚生活](#) >> [服装配饰](#) >> [穿衣搭配吧](#) >> 这些衣服怎么样

[发表新主题](#) | [快速回复](#)

黑色小外套[搭配](#)哪里卖的最便宜?

聪明点比较购物，成就网购达人，快来s
marter.com.cn
www.smarter.com.cn

找鞋包[搭配](#)店 上89178.com

怎样才能挣大钱？ 就开鞋包[搭配](#)店 选择
好项目，成功当老板
www.89178.com

黑色帆布鞋 运动、休闲、增高三..

高贝克帆布内增高鞋，清凉夏日，性感更
注目。货到付款，七天包退换..
www.shihui.tv

共：1页 共0个回复 [首页](#) 1

这些衣服怎么样

【楼主】

http://s.click.taobao.com/t_1?i=qXsqDU0UmpYOSw%3D%3D&p=mm_13744626_0_0&n=11>★★非主流*浪漫淑女R068大圆口雪纺蕾
丝花边短袖上衣★杏色 20元

http://s.click.taobao.com/t_1?i=qXsqDU0UmpYOSw%3D%3D&p=mm_13744626_0_0&n=11>★★非主流*浪漫淑女R068大圆口雪纺蕾
丝花边短袖上衣★杏色 20元

http://s.click.taobao.com/t_1?i=qXjd6Mxd8366tw%3D%3D&p=mm_13744626_0_0&n=11>【韩装经典】最新韩版长款/纯棉短袖圆
领T恤/彩色手绘/卡通人物

http://s.click.taobao.com/t_1?i=qX1eVT8GhfiowA%3D%3D&p=mm_13744626_0_0&n=11>特价初夏新款淑女屋外贸正品韩国韩版热
卖休闲外套短袖T恤米娜

123搜狗导航页面



[新闻](#) [网页](#) [音乐](#) [图片](#) [视频](#) [地图](#) [说吧](#) [更多](#)

搜 索

搜狗网址导航 > 时尚

[把sogou123设为主页](#) | [使用老版](#) | [挑错有礼](#)

时装时尚

搜狐女人-流行服	新浪-服饰潮流	非主流中文网	服饰流行前线	时尚网
YOKA时尚网	都市客	瑞丽女性网	MCC论坛	太平洋时尚女性-
优衣汇	逛街网	女友	中国时装网	时尚论坛
爱表族论坛	搜狗时尚吧	试客网	草莓网	上海服饰
DIY手工俱乐部	1626城中至潮	十字绣论坛	鞋帮	vogue时尚网
米娜国际中文网	中国时尚品牌网	发型屋	香港sasa	海报时尚网
网易女人-服饰	健美先生	化妆品库	e天使	哈发时尚美容美发
时尚起义	凤凰时尚	舍得网	爱物网	易物网
试用网	时尚芭莎	喜试网	中国模特在线	编织人生
品牌世家				

时尚杂志

Z-com	BAZAAR-时尚芭莎	精品购物指南	时尚先生	时尚女报
大都市杂志	时尚.娇点	公主志	时尚新娘	ELLE中文网

浏览记录

[\[清除记录\]](#) [\[停用\]](#)

- 时尚
- 购物
- 搜狗
- 搜狐
- 淘宝网

搜狗广告

[百变达人](#) [免邮费](#) [送韩版饰品](#) 1...

品牌服饰来自韩国拥有10万会员服饰网，韩国明星赞助，每天更新时尚
www.baibiandaren.cn

[休闲服装加盟哪里卖的最便宜?](#)

聪明点比较购物，成就网购达人，快来smarter.com.cn
www.smarter.com.cn

[服装辅料-江浙沪皖服装贸易网](#)

引领服装行业平台，铸就企业品牌，就在江浙沪皖服装贸易网
www.591yf.com

搜索数据介绍

数据体量

**网页量200亿，日更新8亿
访问量达1.8亿（2011）**

搜狗增速显著，环比增长22%

日志数据体量

1.8亿/天 => 60GB

6 month = 180 day

180 * 60GB = 10TB

数据体量

➤ 搜索仅仅这些数据吗？

➤ 网页索引

➤ 广告访问、点击

➤ 推荐服务数据和日志

➤ ...

案例需求

- 解决存储问题！能存储至少1年的数据量。
- 解决计算和访问问题！大部分服务可以使用延时访问，少部分涉及实时请求访问
- 解决扩展问题！可以随时增加、删除服务器来线性扩展服务
- 扩展服务简单！可以随时增加业务、服务，而不影响现有应用

课程使用数据样例

访问时间	用户 ID	[查询词]	URL排名	点击序号	用户点击的 URL
20111230000005	57375476989eeal2893c0c3811607bcf	奇艺高清	1	1	http://www.qiyi.com/
20111230000005	66c5bb7774e31d0a22278249b26bc83a	凡人修仙传	3	1	http://www.booksky.org/BookDetail.as
20111230000007	b97920521c78de70ac38e3713f524b50	本本联盟	1	1	http://www.bblianmeng.com/
20111230000008	6961d0c97fe93701fc9c0d861d096cd9	华南师范大学图书馆	1	1	http://lib.scnu.edu.cn/
20111230000008	f2f5a21c764aebde1e8afcc2871e086f	在线代理	2	1	http://proxyie.cn/
20111230000009	96994a0480e7e1edcaef67b20d8816b7	伟大导演	1	1	http://movie.douban.com/review/11289
20111230000009	698956eb07815439fe5f46e9a4503997	youku	1	1	http://www.youku.com/
20111230000009	599cd26984f72ee68b2b6ebefccf6aed	安徽合肥365房产网	1	1	http://hf.house365.com/
20111230000010	f577230df7b6c532837cd16ab731f874	哈萨克网址大全	1	1	http://www.kz321.com/
20111230000010	285f88780dd0659f5fc8acc7cc4949f2	IQ数码	1	1	http://www.iqshuma.com/
20111230000010	f4ba3f337efb1cc469fcd0b34feff9fb	推荐待机时间长的手机	1	1	http://mobile.zol.com.cn/148
20111230000010	3d1acc7235374d531de1ca885df5e711	满江红	1	1	http://baike.baidu.com/view/6500.htm
20111230000010	dbce4101683913365648eba6a85b6273	光标下载	1	1	http://zhidao.baidu.com/question/386
20111230000011	58e7d0caec23bcb4daa7bbcc4d37f008	张国立的电视剧	2	1	http://tv.sogou.com/vertical/2xc3t6w
20111230000011	a3b83dc38b2bbc35660dffcab4ed9da8	吹暖花开性吧	1	1	http://www.7183.info/
20111230000011	b89952902d7821db37e8999776b32427	怎么骂一个人不带脏字	1	1	http://wenwen.soso.com/z/q13
20111230000011	7c54c43f3a8a0af0951c26d94a57d6c8	百度一下 你就知道	1	1	http://www.baidu.com/
20111230000011	2d6c22c084a501c0b8f7f0a845aefd9f	快播爽吧	5	1	http://www.dy241.com/
20111230000011	11097724dae8b9fdcc60bd6fa4ce4df2	118图库	2	1	http://118123.net/
20111230000012	1d374b57fbbc81aa0cc38e6f4efb88ec	qq老头像	1	1	http://tui.qihoo.com/28302631/articl
20111230000012	76029a8965e815b413cba0b50d2ec2b0	聂磊	1	1	http://baike.baidu.com/view/1941330.htm
20111230000013	22201bdc15845bfb33384efc3a283ef4	cf官网	1	1	http://cf.qq.com/
20111230000013	e0d255845fc9e66b2a25c43a70de4a9a	无饶河, 益慎职 意思	3	1	http://hanyu.iciba.com/wiki/
20111230000013	b89952902d7821db37e8999776b32427	怎么骂一个人不带脏字	2	2	http://zhidao.baidu.com/ques
20111230000013	072fa3643c91b29bd586aff29b402161	暴力破解无线网络密码	2	1	http://download.csdn.net/det
20111230000014	f31f594bd1f3147298bd952ba35de84d	12306.cn	1	1	http://www.12306.cn/
20111230000014	26d2e31d48f527e34c082bc0de591e0e	傲视千雄	1	1	http://web.4399.com/asqx/
20111230000015	66c5bb7774e31d0a22278249b26bc83a	大唐小郎中	3	1	http://www.booksky.org/BookDetail.as
20111230000016	f1b7efe9428b9074f79d3e91ecc6385e	海口双岛幼儿园	3	1	http://www.xugu.net/school/top.asp?i

课程使用数据样例

- 2011年某天数据
 - 4500W记录
 - 2000w独立用户
 - 4.6GB

数据分析需求

- 6大类需求：
- (1) 构建数据仓库 (分月、天)
- (2) 条数统计
 - 数据总条数
 - 非空查询条数
 - 非空非重复条数
 - 独立UID数

数据分析需求

- (3) 关键词分析
 - 查询关键词长度统计
 - 查询频度排名
- (4) UID分析
 - UID查询次数分布
 - UID平均查询次数
 - UID改变查询关键词统计

数据分析需求

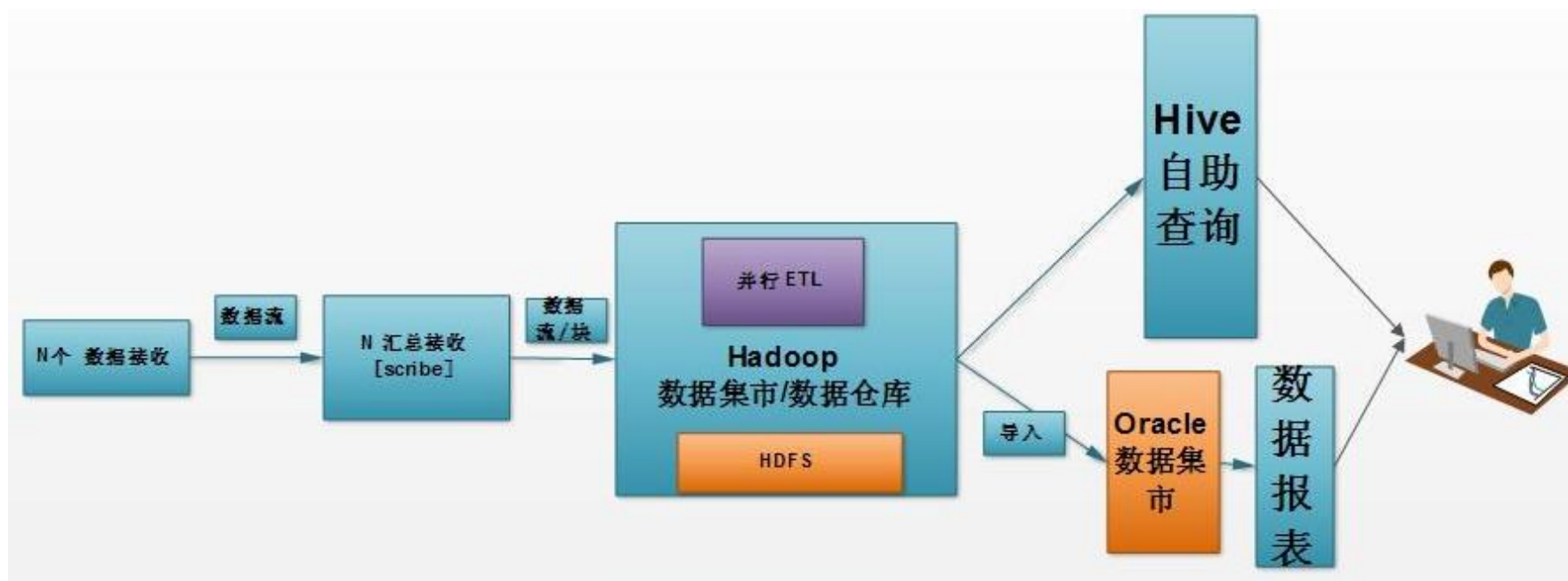
- (5) 用户行为分析
 - 点击次数与Rank之间的关系分析
 - 直接输入URL作为查询词的比例
 - 独立用户行为分析
- (6) 实时数据展示

技术选型与架构

平台和技术架构参考

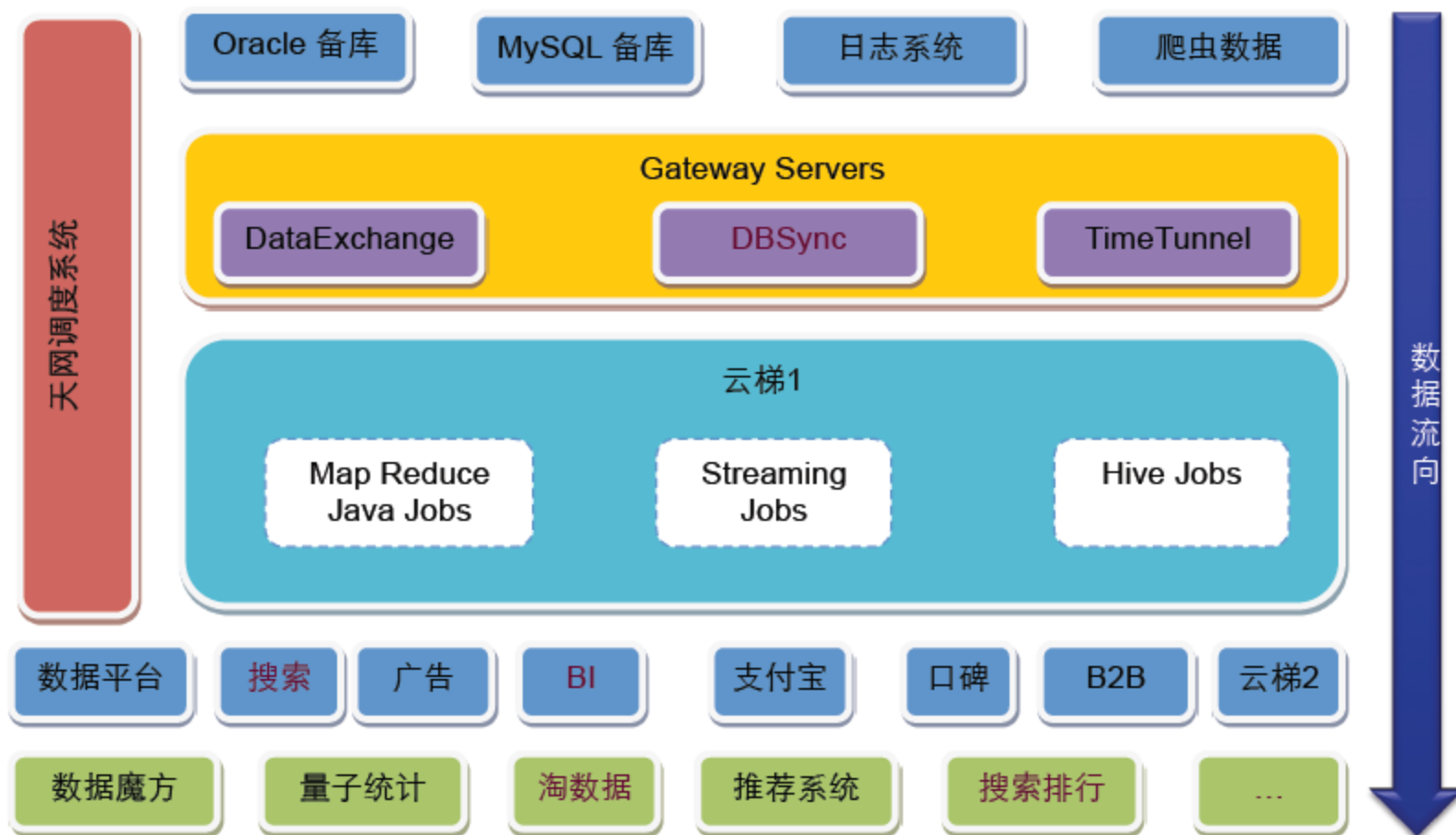
➤ Hadoop后的数据仓库流程

➤ Hive、Pig、MR、工作流



Taobao

系统整体架构



Taobao



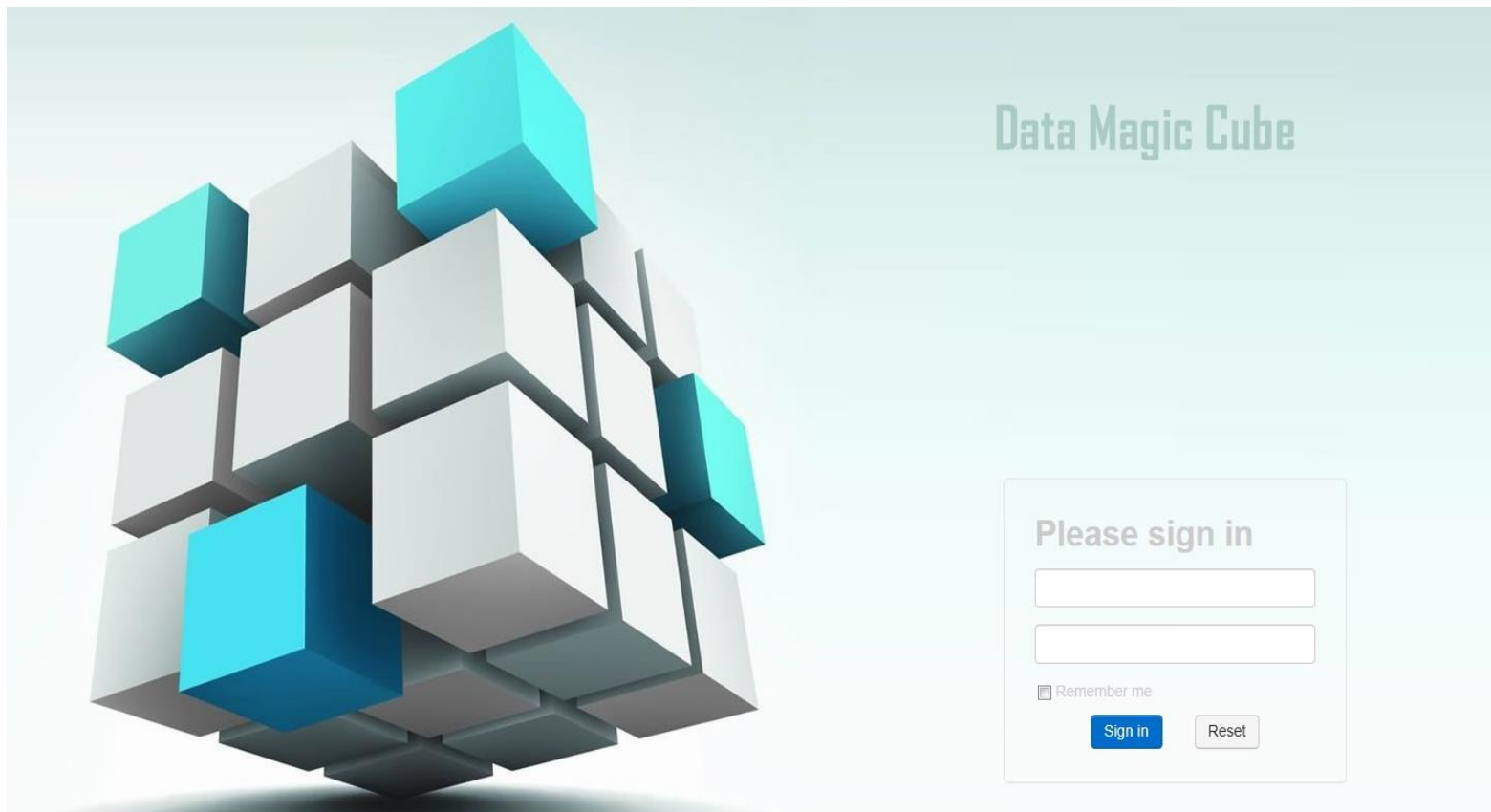
HP



系统架构与核心功能模块

- WEB SERVER->
- 日志采集平台->
- 日志清洗、加载->
- 数据仓库->
- 离线分析系统->
- ETL加载->
- 关系型数据库->
- WEB系统数据展示

最终效果



最终效果

Data Magic Cube

Home

About

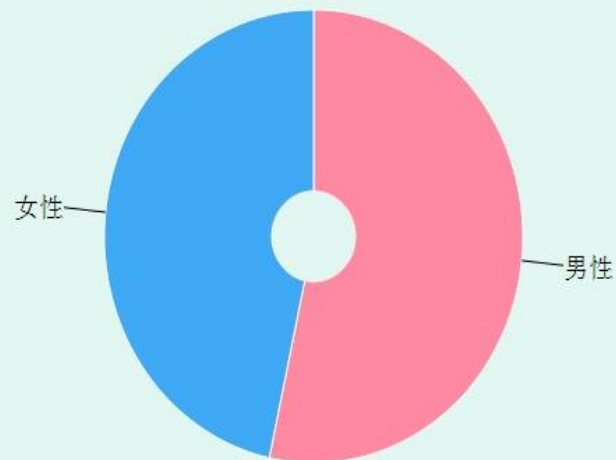
Contact

请输入uid



Gender

■ 男性
■ 女性



最终效果

Data Magic Cube

Home

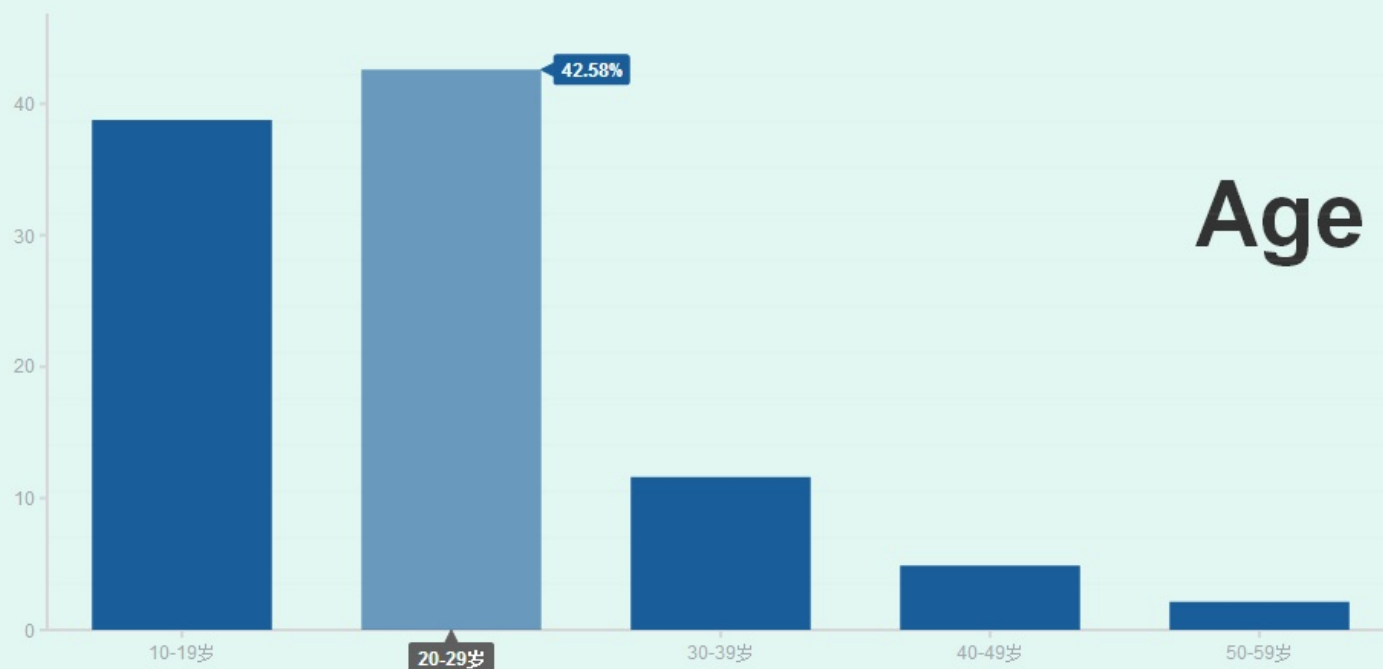
About

Contact

请输入uid



Age



最终效果

Data Magic Cube

Home

About

Contact

请输入uid



Occupation

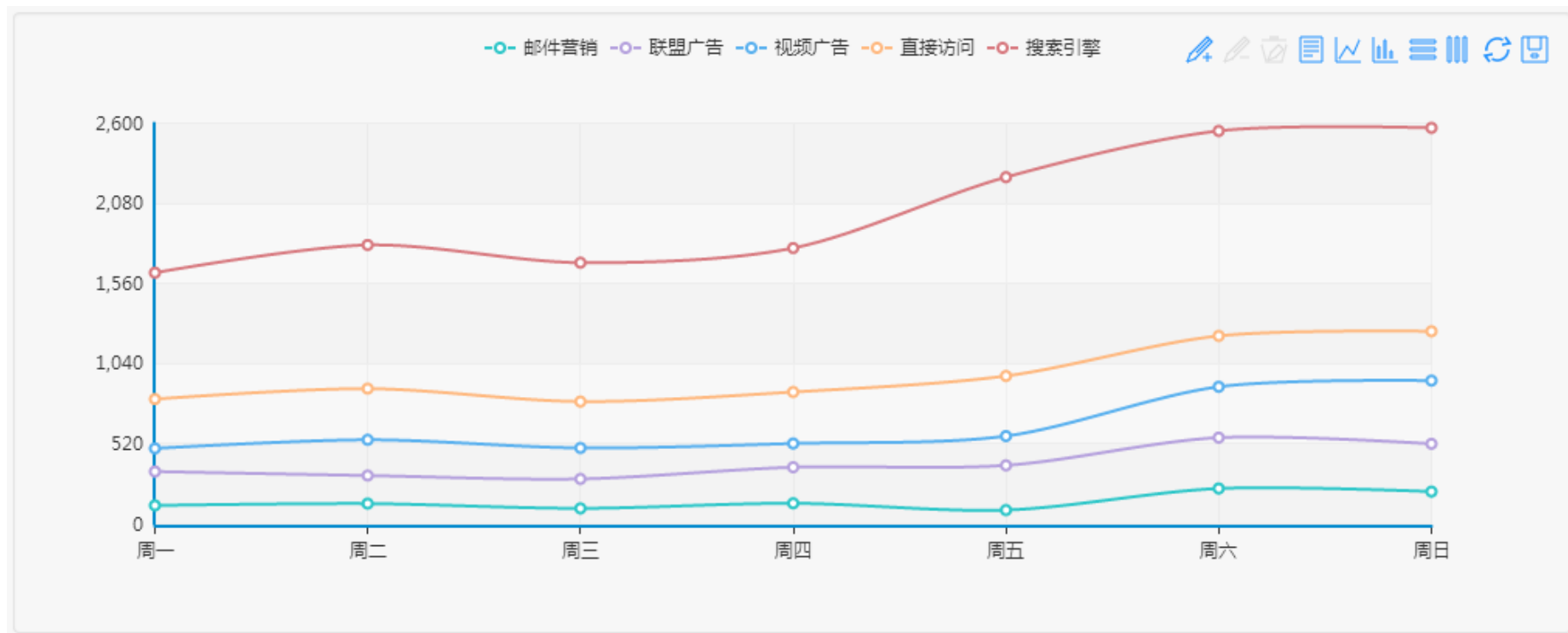
- 1 ~ 10
- 11 ~ 20
- 21 ~ 30
- 30 ~ 39
- 40 ~ 49
- 50 ~ 59



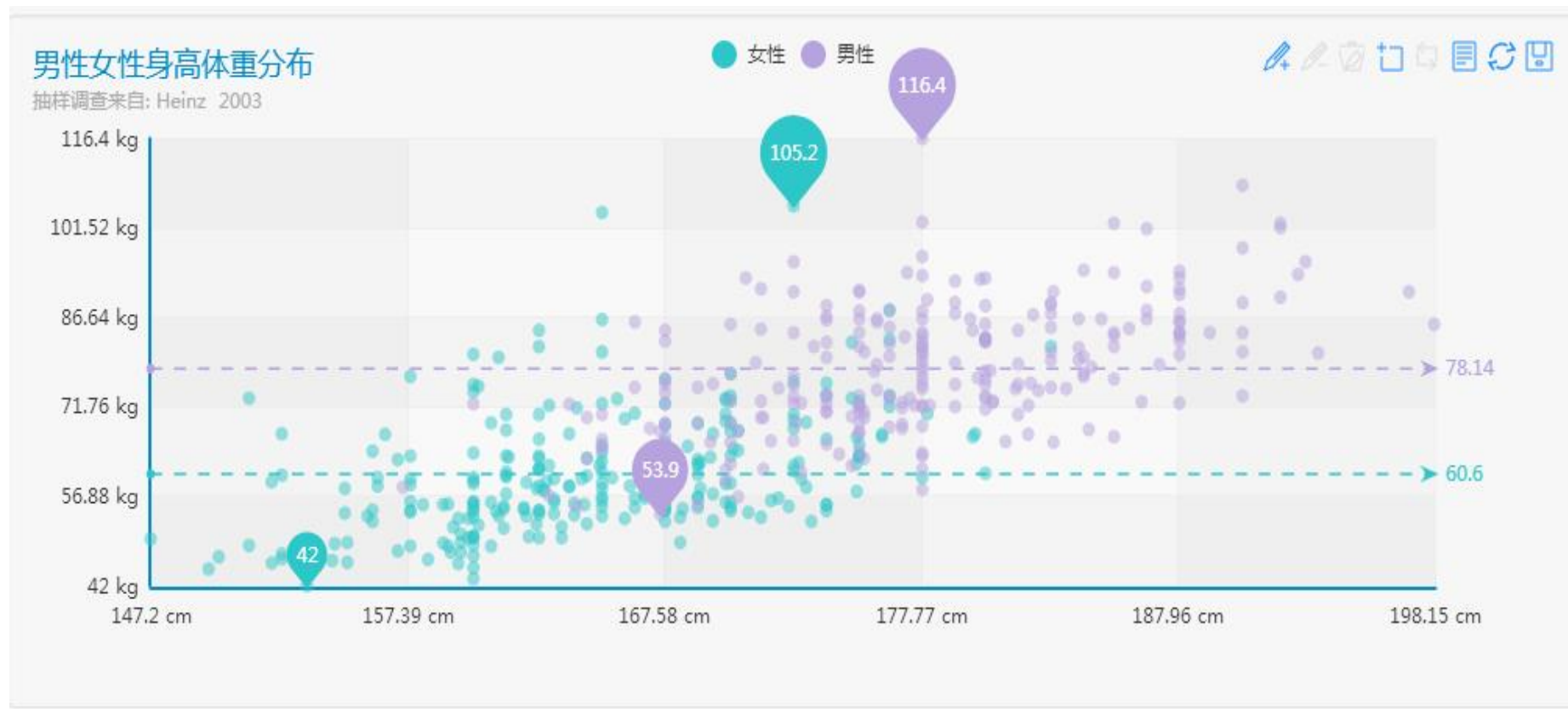
最终效果



最终效果



最终效果



Web Server

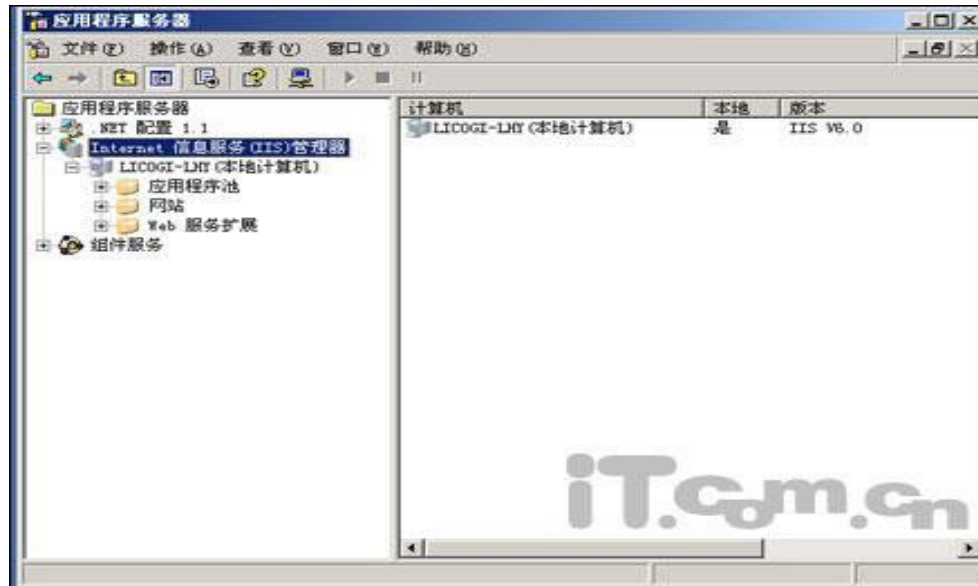
Web server

WWW和Web Server的概念

- WWW(World Wide Web)是Internet上被广泛应用的一种信息服务技术。WWW 采用的是客户/服务器结构，整理和储存各种WWW资源，并响应客户端软件的请求，把所需的信息资源通过浏览器传送给用户。
- 客户指浏览器
- 服务器指Web Server。从功能上讲：WEB服务器是提供网上信息浏览服务，也称为WWW(WORLD WIDE WEB)服务器。

Web server

- 主流的Web server（服务器软件）介绍：
- **Microsoft IIS**



MS IIS

- 微软的Web服务器产品为Internet Information Server (IIS)，IIS 是允许在公共Intranet或Internet上发布信息的Web服务器。例如：微软的官方网站
- IIS提供了一个图形界面的管理工具，称为 Internet 服务管理器，可用于监视配置和控制Internet服务
- IIS包括Web服务器、FTP服务器和SMTP服务器，它使得在网络（包括互联网和局域网）上发布信息成了一件很容易的事。同时，它还提供一个Internet数据库连接器，可以实现对数据库的查询和更新。

IBM WebSphere

创建新服务器和服务配置

创建新服务器和服务配置
选择新服务器的属性。

服务器名 (A) : TASV5

文件夹 (O) : 服务器

服务器类型 (S) :

- WebSphere V5.0
 - EE 测试环境
 - EE 远程服务器
 - 测试环境
 - 远程服务器
 - 远程服务器连接
- WebSphere V4.0
- Apache Tomcat V4.1
- Apache Tomcat V4.0

模板 (T) : 无

描述 : 在本地测试环境上的工作区外部运行 J2EE 项目。

服务器配置类型 (C) : WebSphere v5.0 服务器配置

模板 (M) : 无

描述 : WebSphere V5.0 的服务器配置。

IT.com.cn

< 上一步 (P) > 下一步 (N) > 完成 (F) 取消

IBM WebSphere

- WebSphere软件平台提供了一整套全面的集成电子商务软件解决方案。作为一种基于行业标准的平台，它拥有足够的灵活性，能够适应市场的波动和商业目标的变化。它能够创建、部署、管理、扩展出强大、可移植、与众不同的**电子商务应用**，所有这些内容必要时都可以与现有的传统应用实现集成。以这一稳固的平台为基础，客户可以将不同的IT环境集成在一起，从而能够最大程度地利用现有的投资。

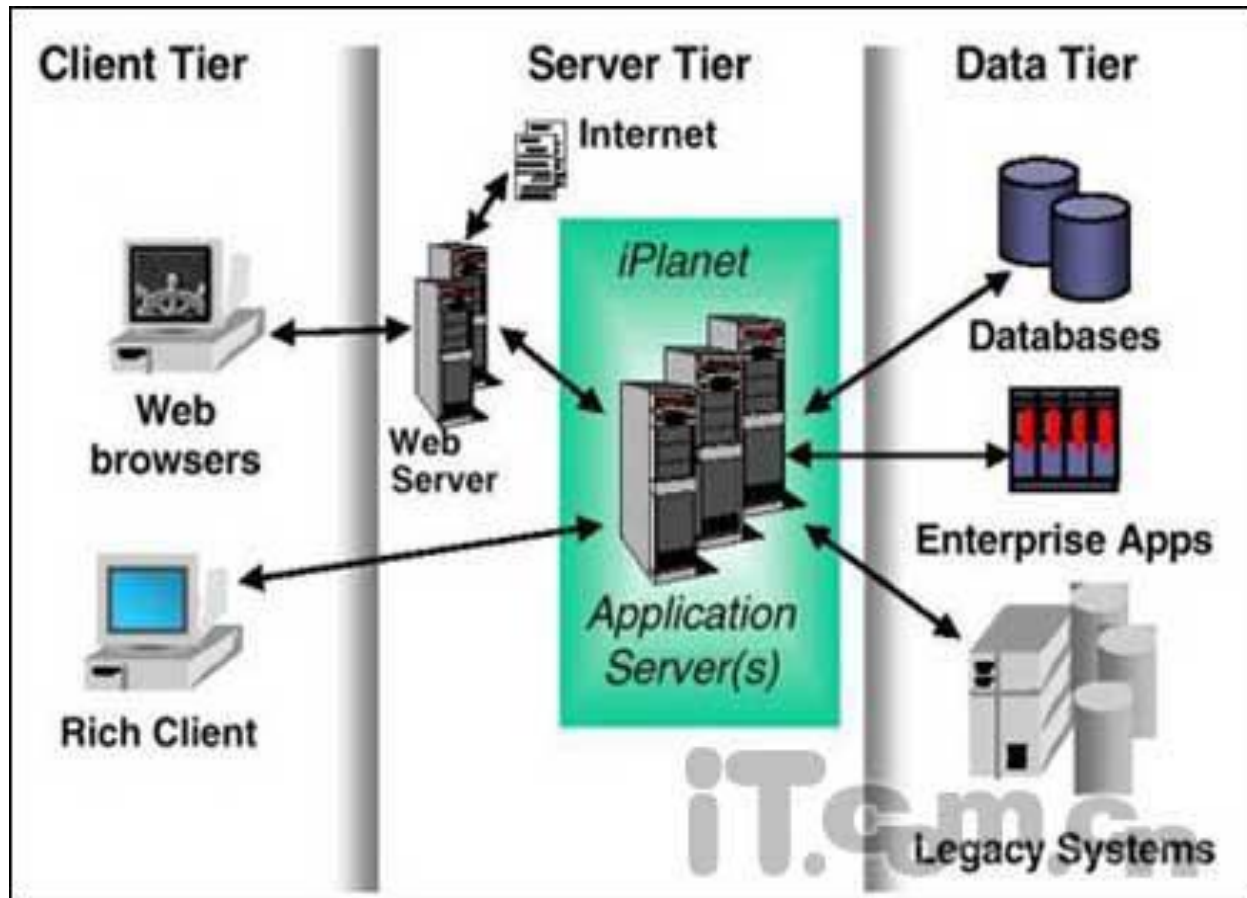
BEA WebLogic



BEA WebLogic

- 具有全面的功能、对开放标准的遵从性、多层架构、支持基于组件的开发，基于 Internet 的企业都选择它来开发、部署最佳的应用。
- BEA WebLogic Server 在使应用服务器成为**企业应用架构**的基础方面继续处于领先地位。
- BEA WebLogic Server 的遵从 J2EE、面向服务的架构，以及丰富的工具集支持，便于实现业务逻辑、数据和表达的分离，提供开发和部署各种业务驱动应用所必需的底层核心功能。

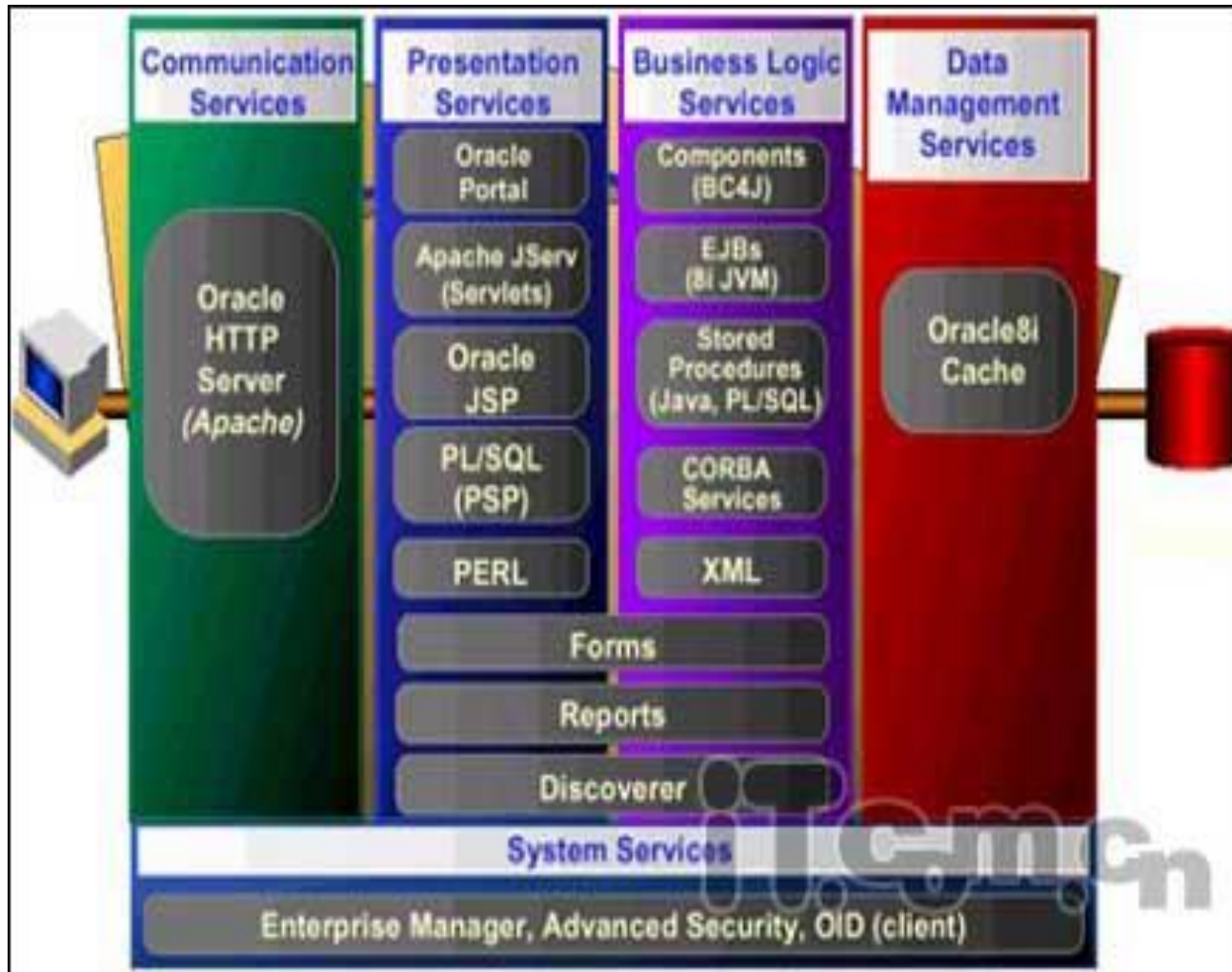
IPlanet Application Server



IPlanet Application Server

- iPlanet Application Server的基本核心服务包括事务监控器、多负载平衡选项、对集群和故障转移全面的支持、集成的XML 解析器和可扩展格式语言转换（XLST）引擎以及对国际化的全面支持。
iPlanet Application Server 企业版所提供的全部特性和功能，并得益于J2EE系统构架，拥有更好的商业**工作流程管理工具和应用集成功能**。

Oracle IAS



Oracle IAS

- Oracle Internet Application Server 最大的优势是其集成性和通用性
- 集成性方面支持Web浏览器、胖客户端和手持设备
- 在通用性方面，Oracle iAS支持各种业界标准，包括 JavaBeans、CORBA、Servlets以及XML标准等

Apache



Apache

- Apache是自由软件，所以不断有人来为它开发新的功能、新的特性、修改原来的缺陷。Apache的特点是简单、速度快、性能稳定，并可做代理服务器来使用。本来它只用于小型或试验Internet网络，后来逐步扩充到各种Unix系统中，尤其对Linux的支持相当完美。

Nginx

- Nginx (“engine x”) 是俄罗斯人Igor Sysoev(塞索耶夫)编写的一款高性能的 HTTP 和反向代理服务器。
- Nginx 已经在俄罗斯最大的门户网站—— Rambler Media (www.rambler.ru) 上运行了4年时间，同时俄罗斯超过20%的虚拟主机平台采用Nginx作为反向代理服务器。
- 在国内，已经有新浪博客、新浪播客、搜狐通行证、网易新闻、网易博客、金山逍遥网等多家网站、频道使用 Nginx 服务器。

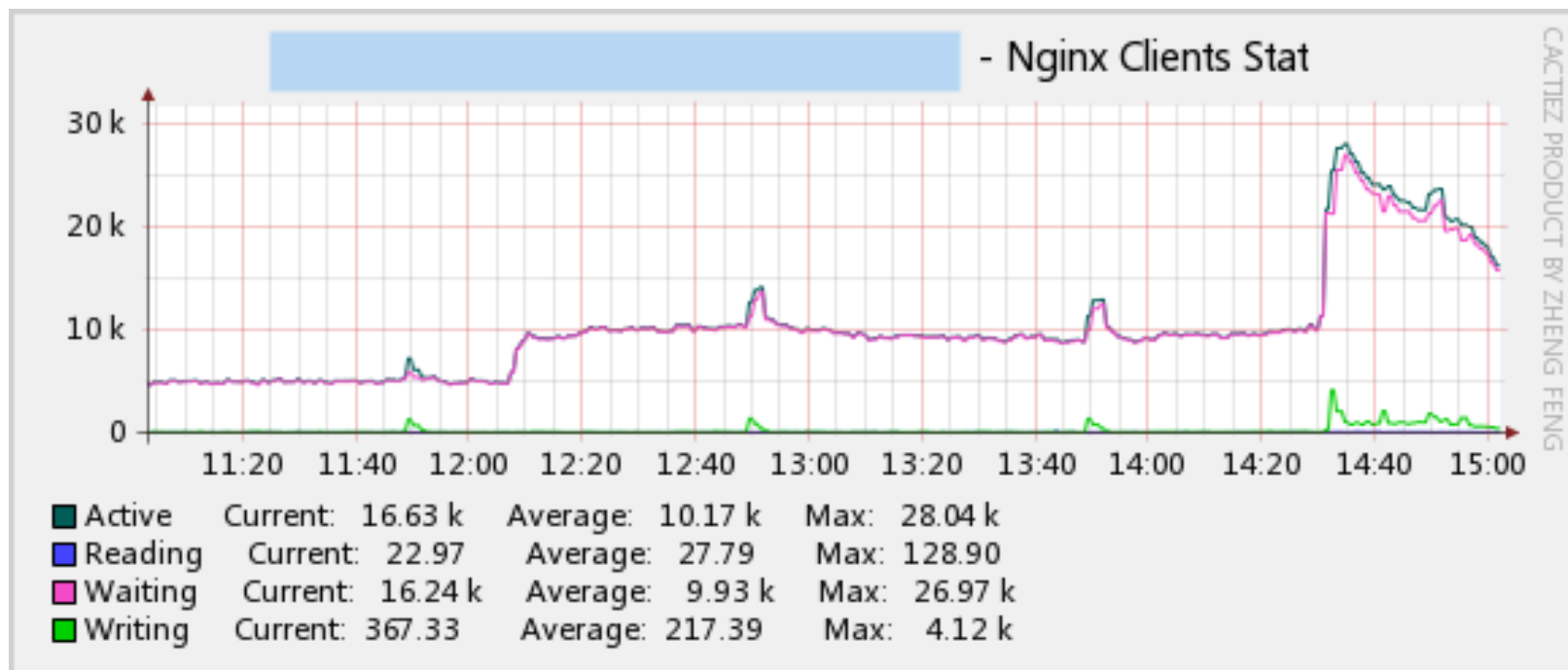
Nginx的优点①

- **1、高并发连接**：官方测试能够支撑5万并发连接，在实际生产环境中跑到2 ~ 3万并发连接数。
- **2、内存消耗少**：在3万并发连接下，开启的10个Nginx 进程才消耗150M内存（ $15\text{M} \times 10 = 150\text{M}$ ）。
- **3、配置文件非常简单**：风格跟程序一样通俗易懂。
- **4、成本低廉**：Nginx为开源软件，可以免费使用。而购买F5 BIG-IP、NetScaler等硬件负载均衡交换机则需要十多万至几十万人民币。

Nginx的优点②

- 5、支持Rewrite重写规则：能够根据域名、URL的不同，将 HTTP 请求分到不同的后端服务器群组。
- 6、内置的健康检查功能：如果 Nginx Proxy 后端的某台 Web 服务器宕机了，不会影响前端访问。
- 7、节省带宽：支持 GZIP 压缩，可以添加浏览器本地缓存的 Header 头。
- 8、稳定性高：用于反向代理，宕机的概率微乎其微。

Nginx单点性能



➤ 2009-09-03 14:30，金山游戏《剑侠情缘网络版3》临时维护1小时，大量玩家上官网，论坛、评论、客服等动态应用Nginx服务器集群，每台服务器的Nginx活动连接数达到2.8万，Nginx生产环境相当高的并发值。

Nginx的主要应用类别

- 1、使用 Nginx 结合FastCGI运行 PHP、JSP 、 Perl 等程序
- 2、使用 Nginx 作反向代理、负载均衡、规则过滤
- 3、使用 Nginx 运行静态HTML页、图片
- 4、Nginx与其他新技术的结合应用

总结与提问



开源日志收集系统

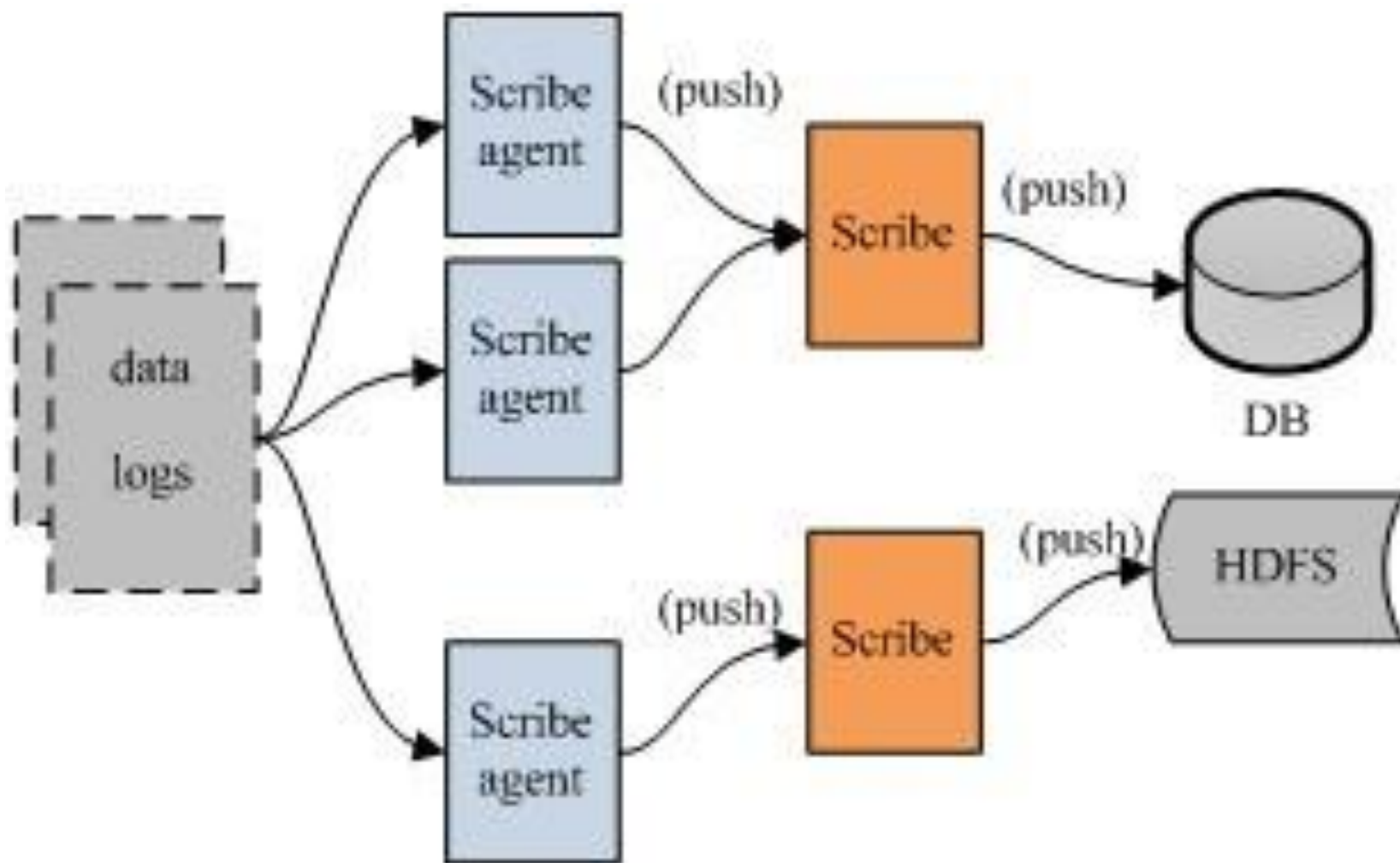
Scribe

➤ Scribe是facebook开源的日志收集系统，在facebook内部已经得到大量的应用。它能够从各种日志源上收集日志，存储到一个中央存储系统（可以是NFS，分布式文件系统等）上，以便于进行集中统计分析处理。它为日志的“分布式收集，统一处理”提供了一个可扩展的，高容错的方案。

Scribe

➤最重要的特点是容错性好。当后端的存储系统crash时，scribe会将数据写到本地磁盘上，当存储系统恢复正常后，scribe将日志重新加载到存储系统中。

Scribe数据流



Scribe架构

➤scribe的架构比较简单，主要包括三部分，分别为scribe agent，scribe和存储系统。

➤(1) scribe agent

➤scribe agent实际上是一个thrift client。向scribe发送数据的唯一方法是使用thrift client，scribe内部定义了一个thrift接口，用户使用该接口将数据发送给server

Scribe架构

➤(2) scribe

➤scribe接收到thrift client发送过来的数据，根据配置文件，将不同topic的数据发送给不同的对象。scribe提供了各种各样的store，如file，HDFS等，scribe可将数据加载到这些store中。

➤(3) 存储系统

➤存储系统实际上就是scribe中的store，当前scribe支持非常多的store，包括file（文件），buffer（双层存储，一个主储存，一个副储存），network（另一个scribe服务器），bucket（包含多个store，通过hash的将数据存到不同store中），null(忽略数据)，thriftfile（写到一个Thrift TFileTransport文件中）和multi（把数据同时存放到不同store中）。

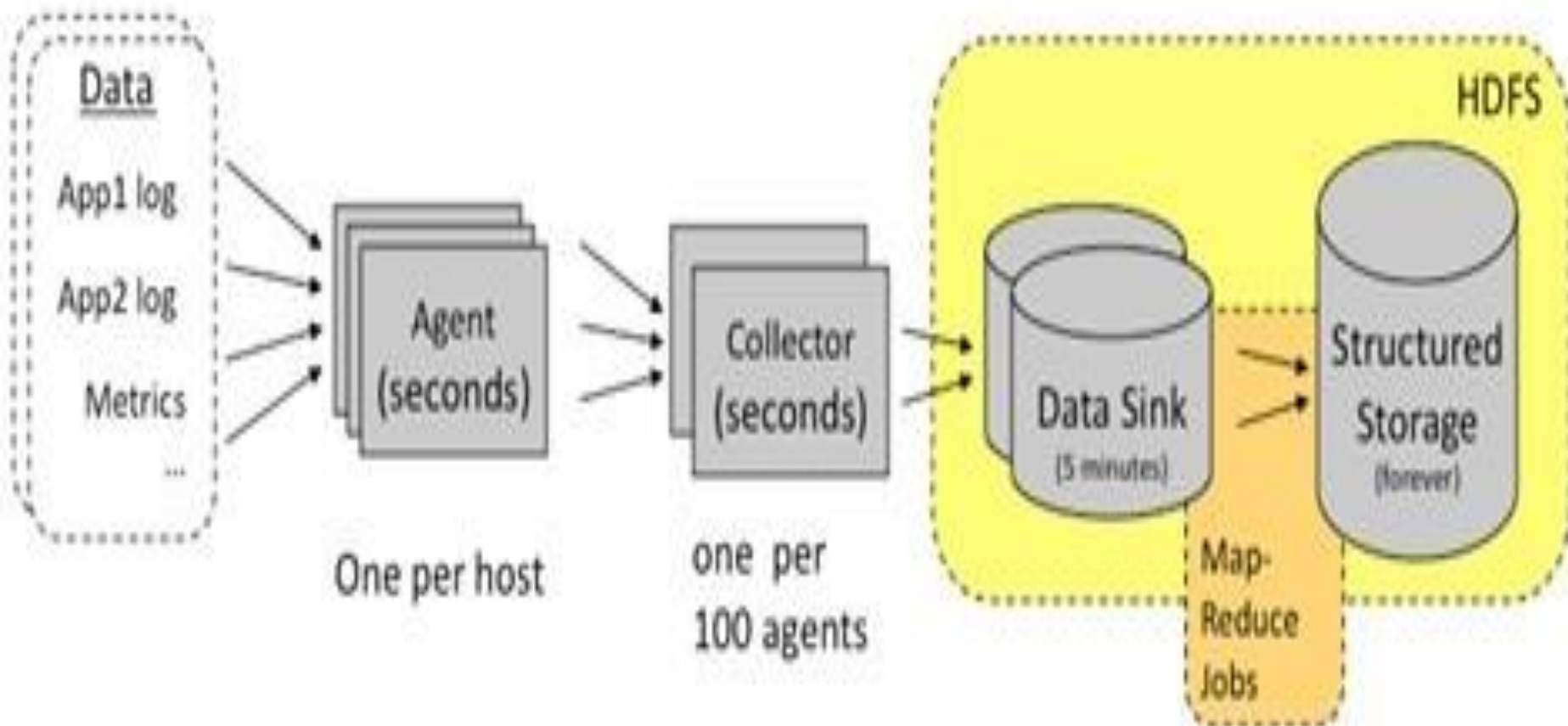
Apache Chukwa

➤chukwa是一个非常新的开源项目，由于其属于hadoop系列产品，因而使用了很多hadoop的组件（用HDFS存储，用mapreduce处理数据），它提供了很多模块以支持hadoop集群日志分析。

➤需求：

- (1) 灵活的，动态可控的数据源
- (2) 高性能，高可扩展的存储系统
- (3) 合适的框架，用于对收集到的大规模数据进行分析

Apache Chukwa数据流



Chukwa架构

- Chukwa中主要有3种角色，分别为：adaptor，agent，collector。
- (1) Adaptor 数据源
 - 可封装其他数据源，如file，unix命令行工具等
 - 目前可用的数据源有：hadoop logs，应用程序度量数据，系统参数数据（如linux cpu使用流率）。
- (2) HDFS 存储系统
 - Chukwa采用了HDFS作为存储系统。HDFS的设计初衷是支持大文件存储和小并发高速写的应用场景，而日志系统的特点恰好相反，它需支持高并发低速率的写和大量小文件的存储。需要注意的是，直接写到HDFS上的小文件是不可见的，直到关闭文件，另外，HDFS不支持文件重新打开。

Chukwa架构

➤(3) Collector和Agent

➤为了克服(2)中的问题，增加了agent和collector阶段

➤Agent的作用：给adaptor提供各种服务，包括：启动和关闭adaptor，将数据通过HTTP传递给Collector；定期记录adaptor状态，以便crash后恢复。

➤Collector的作用：对多个数据源发过来的数据进行合并，然后加载到HDFS中；隐藏HDFS实现的细节，如，HDFS版本更换后，只需修改collector即可。

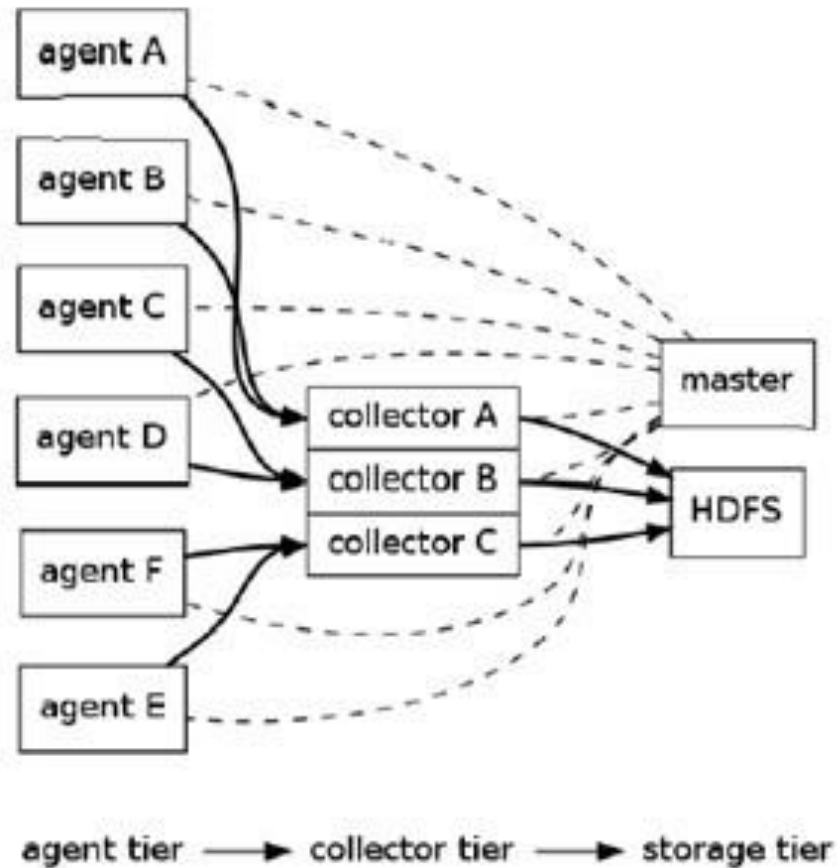
➤(4) Demux和achieving

➤直接支持利用MapReduce处理数据。它内置了两个mapreduce作业，分别用于获取data和将data转化为结构化的log。存储到data store（可以是数据库或者HDFS等）中。

Flume

➤ Flume是cloudera于2009年7月开源的日志系统。它内置的各种组件非常齐全，用户几乎不必进行任何额外开发即可使用。

Flume



Flume架构

➤正如前面提到的，Flume采用了分层架构，由三层组成，分别为agent，collector和storage。其中，agent和collector均由两部分组成：source和sink，source是数据来源，sink是数据去向。

Flume架构

➤(1) agent

➤agent的作用是将数据源的数据发送给collector ,
Flume自带了很多直接可用的数据源 (source)

➤(2) collector

➤collector的作用是将多个agent的数据汇总后 , 加载到storage中。它的source和sink与agent类似。

Flume架构

➤(3) storage

➤storage是存储系统，可以是一个普通file，也可以是HDFS，HIVE，HBase等。

日志收集系统对比

	scribe	Chukwa	Kafka	Flume
公司	facebook	apache/yahoo	LinkedIn	Cloudera
开源时间	2008 年 10 月	2009 年 11 月	2010 年 12 月	2009 年 7 月
实现语言	C/C++	JAVA	SCALA	JAVA
框架	push/push	push/push	push/pull	push/push
容错性	collector 和 store 之间有容错机制。而 agent 和 collector 之间的容错需用户自己实现	Agent 定期记录已送给 collector 的数据偏移量，一旦出现故障后，可根据偏移量继续发送数据。	Agent 可用通过 collector 自动识别机制获取可用 collector。store 自己保存已经获取数据的偏移量，一旦 collector 出现故障，可根据偏移量继续获取数据。	Agent 和 collector，collector 和 store 之间均有容错机制，且提供了三种级别的可靠性保证。
负载均衡	无	无	使用 zookeeper	使用 zookeeper
可扩展性	好	好	好	好
agent	Thrift client，需自己实现	自带一些 agent，如获取 hadoop logs 的 agent	用户需根据 Kafka 提供的 low-level 和 high-level API 自己实现。	提供了各种非常丰富的 agent
collector	实际上是一个 thrift server	--	使用了 sendfile，zero-copy 等技术提高性能	系统提供了很多 collector，直接使用。
store	直接支持 HDFS	直接支持 HDFS	直接支持 HDFS	直接支持 HDFS
总体评价	设计简单，易于使用，但容错和负载均衡方面不够好，且资料较少。	属于 hadoop 系列产品，直接支持 Hadoop，目前版本升级比较快，但还有待完善。	设计架构（push/pull）非常巧妙，适合异构集群，但产品较新，其稳定性有待验证。	非常优秀