



# Hadoop应用开发实战案例 第15周

**【声明】** 本视频和幻灯片为炼数成金网络课程的教学资料，所有资料只能在课程内使用，不得在课程以外范围散播，违者将可能被追究法律和经济责任。

课程详情访问炼数成金培训网站

<http://edu.dataguru.cn>

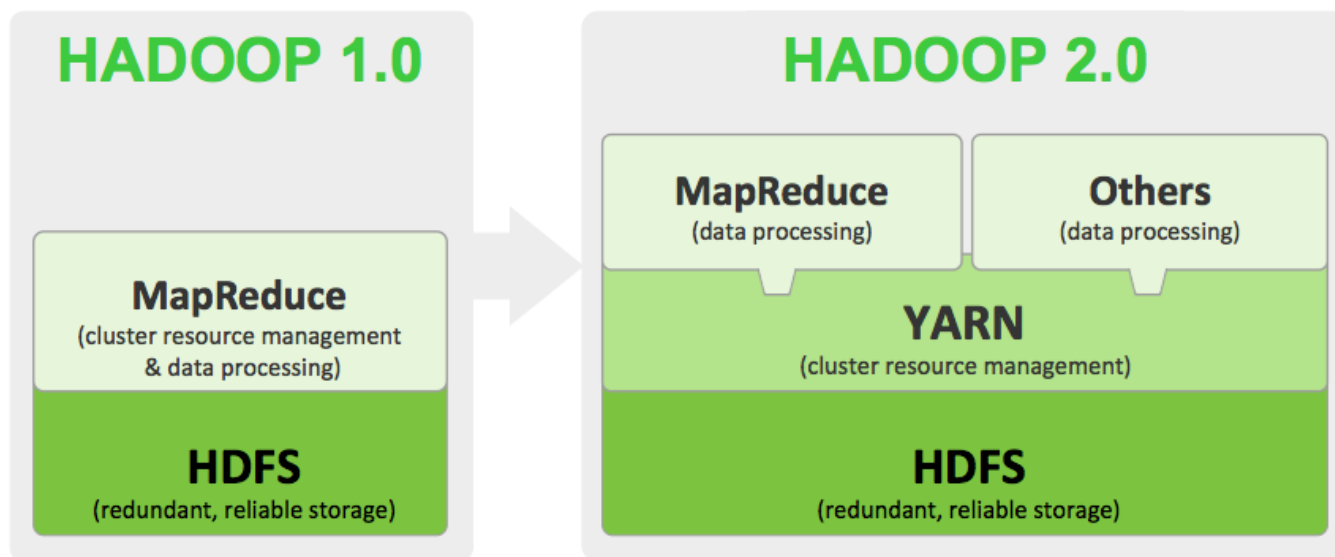
- Hadoop新一代计算平台YARN
- 新一代快速计算平台Spark及其生态圈
- Mahout告别Map-Reduce
- 阿里巴巴抛弃云梯（Hadoop集群）

# 第一代Map-Reduce框架的缺点

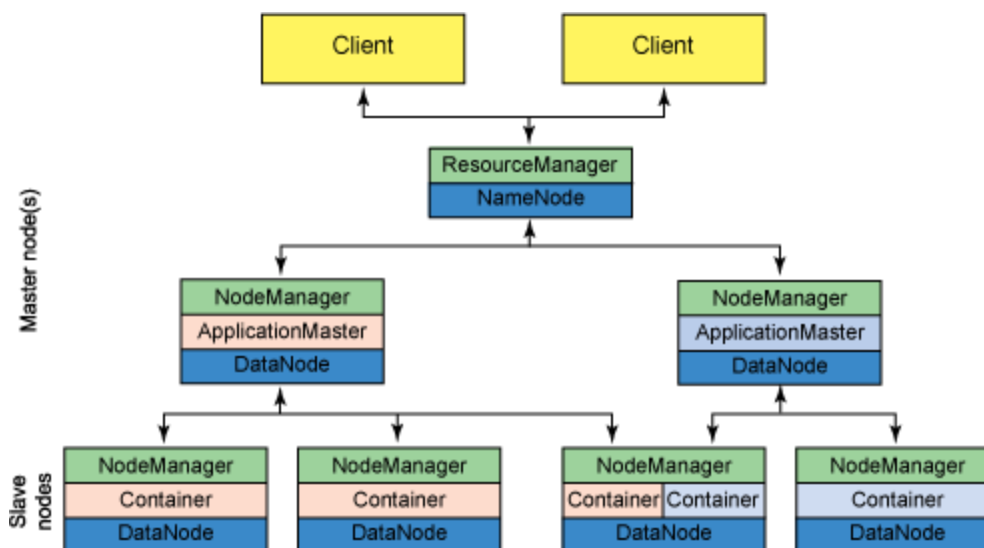
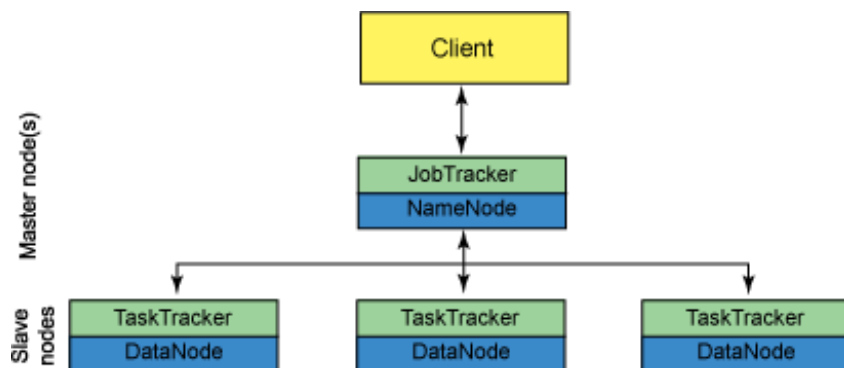
- Jobtracker单点，容易导致崩溃，节点较多时造成性能瓶颈
- 作业分配基于槽位（slot），分配粒度太粗
- Jobtracker和Tasktracker多次来回方能启动作业，导致小作业不能及时完成
- 计算框架单一，Map-Reduce擅长日志分析，但却有大量的机器学习算法需要反复循环迭代，还有像图计算，可能涉及数据不多，但却要在内存产生大量中间数据和超大计算量，这些都不适合使用M-R框架，但Hadoop 1.x却无法支持流式数据库，基于内存的计算这些框架

# 新一代计算平台YARN

- Yet Another Resource Negotiator
- Hadoop 0.23开始引入
- 学习Mesos
- 弹性平台，可以同时支持Map-Reduce，Storm，Spark，MPI等多种流行计算模型



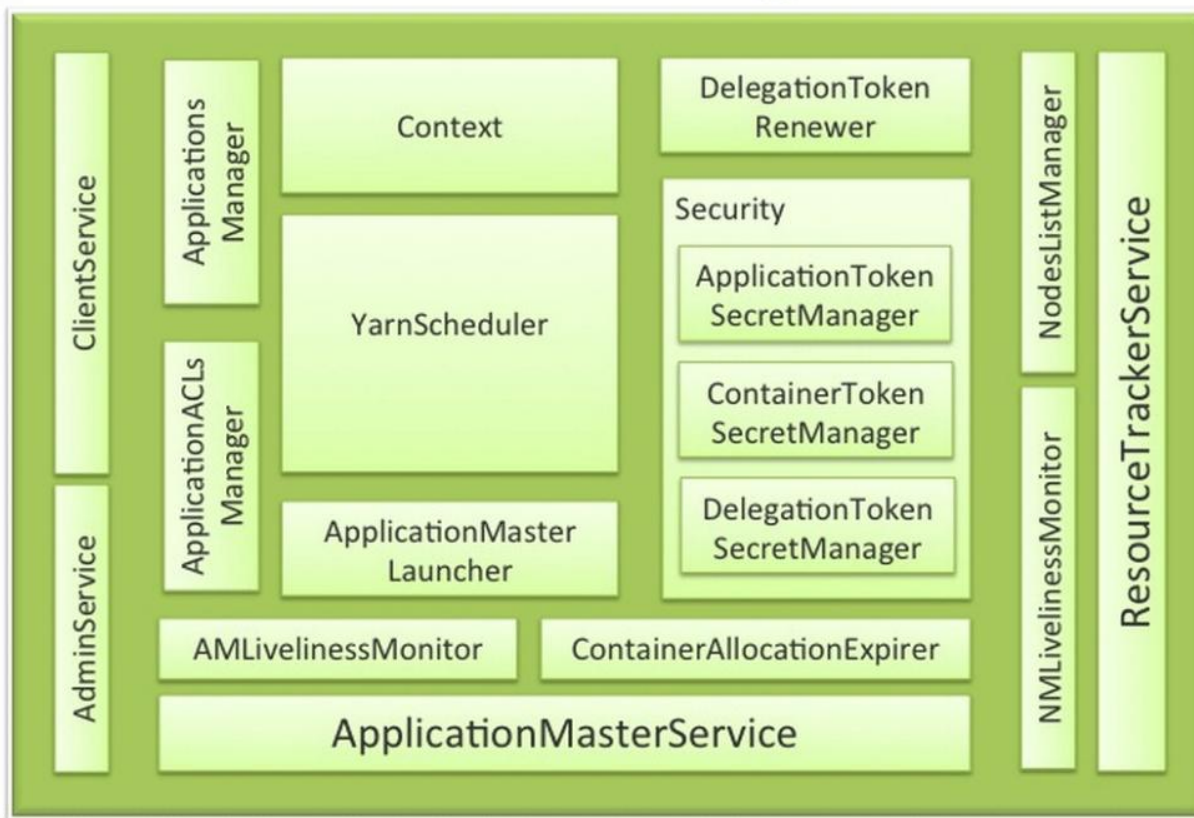
# YARN与MRv1比较



- 控制整个集群并管理应用程序向基础计算资源的分配。
- 将各个资源部分（计算、内存、带宽等）精心安排给基础 NodeManager（YARN 的每节点代理）。
- ResourceManager 还与 ApplicationMaster 一起分配资源，与 NodeManager 一起启动和监视它们的基础应用程序。
- ApplicationMaster 承担了以前的 TaskTracker 的一些角色，ResourceManager 承担了 JobTracker 的角色。

<http://bbs.itcast.cn/thread-17905-1-1.html>

## ResourceManager



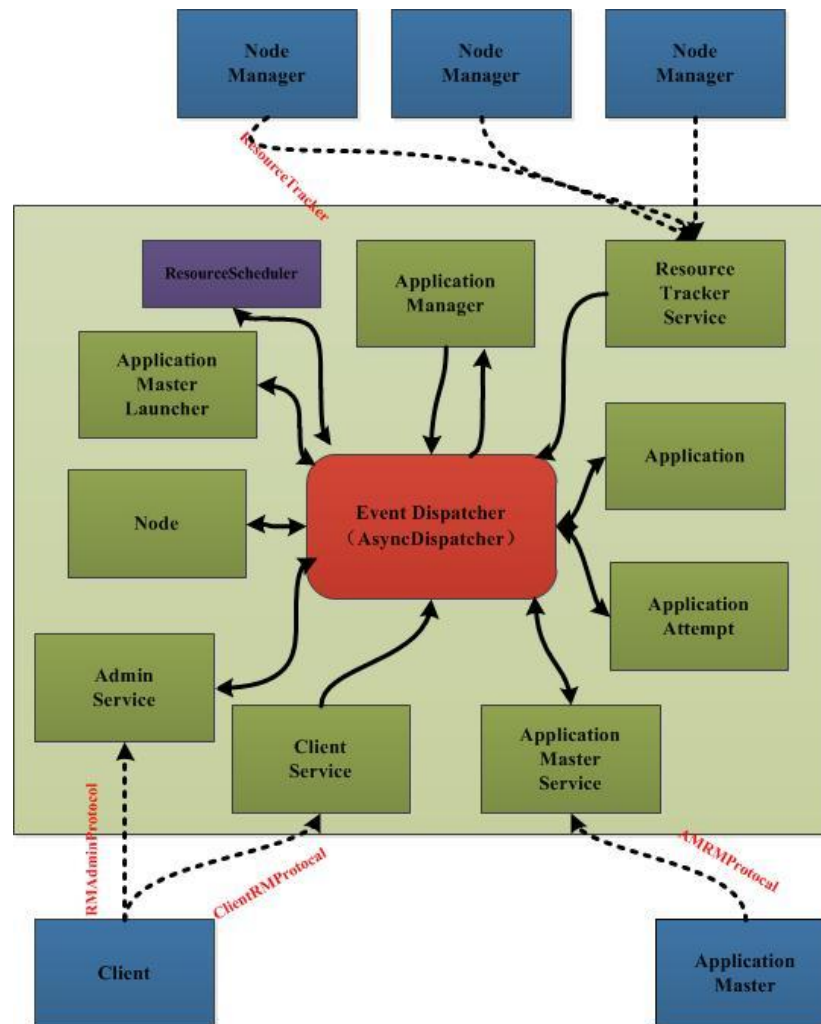


- ApplicationMaster 管理一个在 YARN 内运行的应用程序的每个实例。
- 负责协调来自 ResourceManager 的资源，并通过 NodeManager 监视容器的执行和资源使用（CPU、内存等的资源分配）。请注意，尽管目前的资源更加传统（CPU 核心、内存），但未来会带来基于手头任务的新资源类型（比如图形处理单元或专用处理设备）。
- 从 YARN 角度讲，ApplicationMaster 是用户代码，因此存在潜在的安全问题。YARN 假设 ApplicationMaster 存在错误或者甚至是恶意的，因此将它们当作无特权的代码对待。

- NodeManager 管理一个 YARN 集群中的每个节点。
- 提供针对集群中每个节点的服务，从监督对一个容器的终生管理到监视资源和跟踪节点健康。MRv1 通过插槽管理 Map 和 Reduce 任务的执行，而 NodeManager 管理抽象容器，这些容器代表着可供一个特定应用程序使用的针对每个节点的资源。
- YARN 继续使用 HDFS 层。它的主要 NameNode 用于元数据服务，而 DataNode 用于分散在一个集群中的复制存储服务。

- 要使用一个 YARN 集群，首先需要来自包含一个应用程序的客户的请求。
- ResourceManager 协商一个容器的必要资源，启动一个 ApplicationMaster 来表示已提交的应用程序。
- 通过使用一个资源请求协议，ApplicationMaster 协商每个节点上供应用程序使用的资源容器。
- 执行应用程序时，ApplicationMaster 监视容器直到完成。
- 当应用程序完成时，ApplicationMaster 从 ResourceManager 注销其容器，执行周期就完成了。

# YARN通讯协议



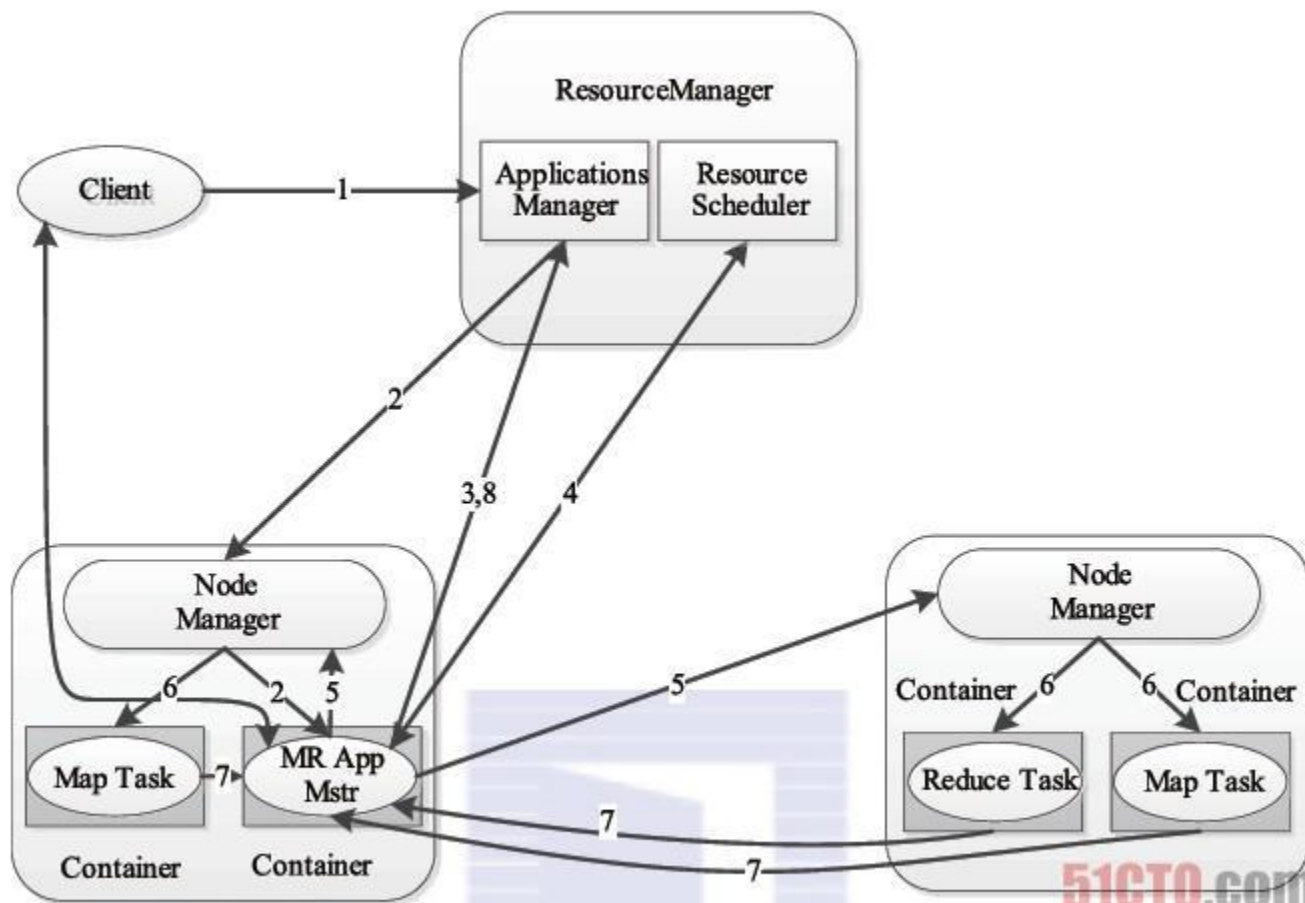
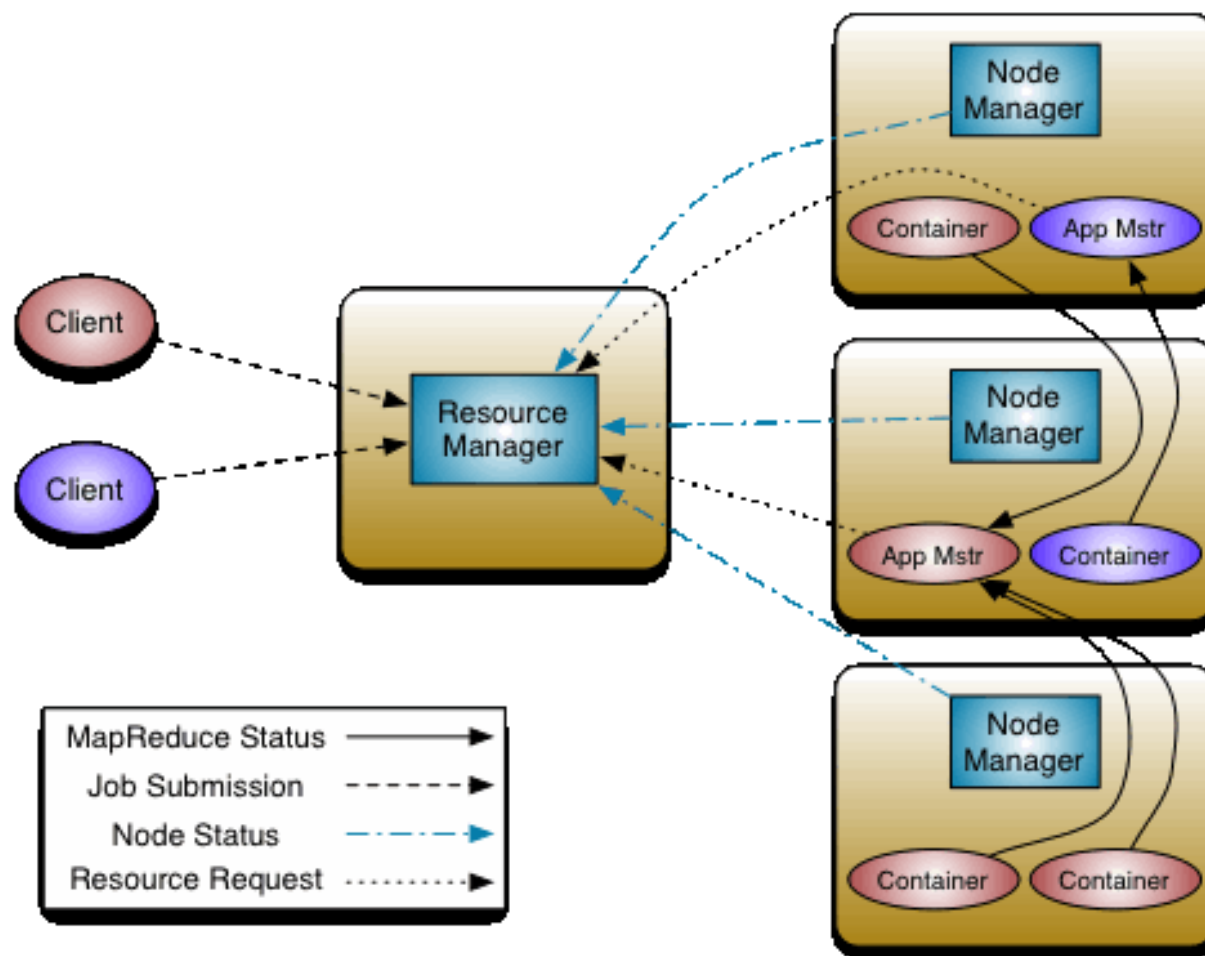


图 2-11 Apache YARN 的工作流程

51CTO.com  
技术成就梦想

# YARN工作流程



- [http://hadoop.apache.org/docs/r2.3.0/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduce Compatibility Hadoop1 Hadoop2.html](http://hadoop.apache.org/docs/r2.3.0/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduce%20Compatibility%20Hadoop1%20Hadoop2.html)
- 对使用老**mapred** API的应用提供二进制兼容性
- 对使用**mapreduce** API的应用提供源码兼容性，也就是需要重新编译执行

# 新一代计算框架Spark



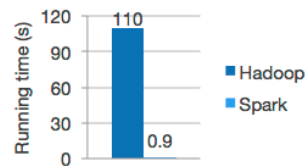
← → ↻ spark.apache.org

Apache Spark™ is a fast and general engine for large-scale data processing.

## Speed

Run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk.

Spark has an advanced DAG execution engine that supports cyclic data flow and in-memory computing.



Logistic regression in Hadoop and Spark

## Ease of Use

Write applications quickly in Java, Scala or Python.

Spark offers over 80 high-level operators that make it easy to build parallel apps. And you can use it *interactively* from the Scala and Python shells.

```
file = spark.textFile("hdfs://...")

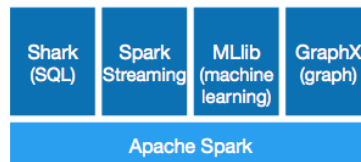
file.flatMap(lambda line: line.split())
      .map(lambda word: (word, 1))
      .reduceByKey(lambda a, b: a+b)
```

Word count in Spark's Python API

## Generality

Combine SQL, streaming, and complex analytics.

Spark powers a stack of high-level tools including [Shark](#) for SQL, [MLlib](#) for machine learning, [GraphX](#), and [Spark Streaming](#). You can combine these frameworks seamlessly in the same application.



### Latest News

Spark Summit agenda posted (May 11, 2014)

Spark 0.9.1 released (Apr 09, 2014)

Submissions and registration open for Spark Summit 2014 (Mar 20, 2014)

Spark becomes top-level Apache project (Feb 27, 2014)

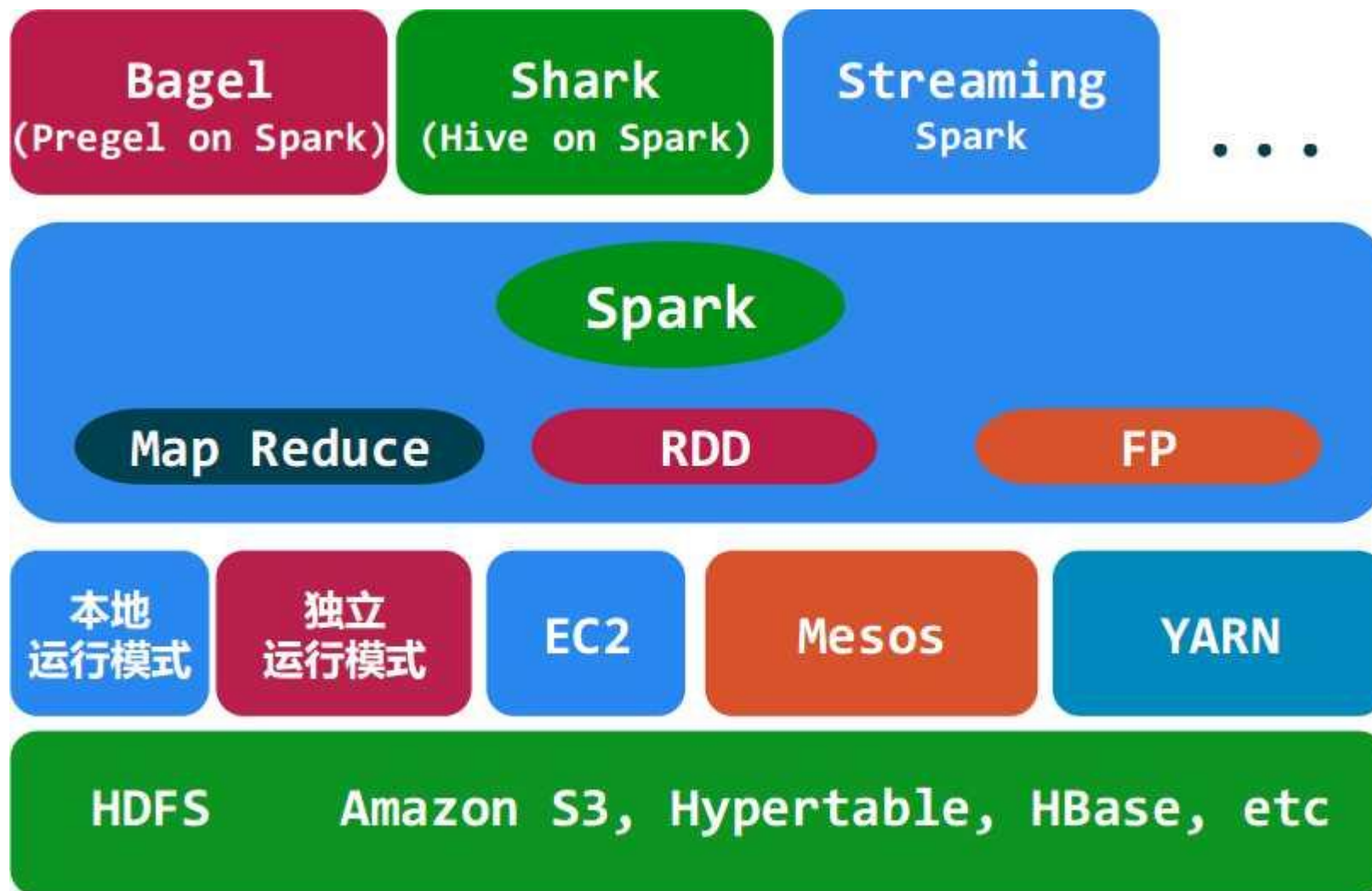
[Archive](#)

[Download Spark](#)

### Related Projects:

- [Shark](#) (SQL)
- [Spark Streaming](#)
- [MLlib](#) (machine learning)
- [GraphX](#) (graph)



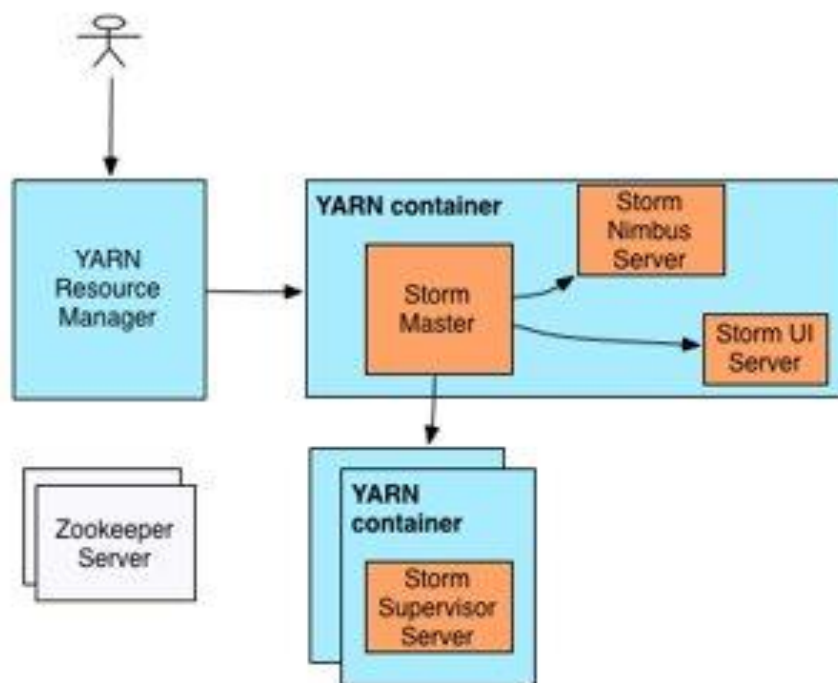


- 淘宝使用情况：

[http://wenku.baidu.com/link?url=g2JNG0EFb1C9dD7gFKHq4v5Keap0mQhJtw0wiqF5oBQI0f0RNnXrtIbUj-zbdqEDhZuR5P7ImE8TrNFPlwTQ3kgSg3FUI57Ze\\_oTB9bpe7](http://wenku.baidu.com/link?url=g2JNG0EFb1C9dD7gFKHq4v5Keap0mQhJtw0wiqF5oBQI0f0RNnXrtIbUj-zbdqEDhZuR5P7ImE8TrNFPlwTQ3kgSg3FUI57Ze_oTB9bpe7)

# Storm on YARN

- 部署：<http://blog.csdn.net/jiushuai/article/details/18729367>
- 源代码：<https://github.com/yahoo/storm-yarn>



- Message Passing Interface
- MPI是一个并行函数库标准，是应用程序对消息传递的需求
- MPICH2是MPI的开源实现



[MPICH2 Home Page](#)  
[Download](#)  
[Documentation](#)  
[License](#)  
[Information for Developers](#)  
[Bug Reports](#)

## MPICH2

MPICH2 is an implementation of the [Message-Passing Interface](#) (MPI). The goals of MPICH2 are to provide an MPI implementation for important platforms, including clusters, SMPs, and massively parallel processors. It also provides a vehicle for MPI implementation research and for developing new and better parallel programming environments.

### Download MPICH2




The current version of MPICH2 is 1.0, released on November 9, 2004. MPICH2 is distributed as source (with an open-source, freely available [license](#)). It has been extensively tested on several platforms, including Linux (on IA32 and IA64), and Windows. See the RELEASE\_NOTES file in the distribution for more details. If you have any questions, comments, or difficulties, please contact us at [mpich2-maint@mcs.anl.gov](mailto:mpich2-maint@mcs.anl.gov).

MPICH2 is provided in source form as a single, gzip'ed tar file. This is a unified source distribution, and may be used for both UNIX and Microsoft Windows. Binary distributions for Microsoft Windows are also provided.

Platform	Download			Size	Version
All (source)	mpich2-1.0.tar.gz	<a href="#">http</a>	<a href="#">ftp</a>	12.2MB	1.0
Win32 IA32	mpich2-1.0-1-win32-ia32.msi	<a href="#">http</a>	<a href="#">ftp</a>	1.7MB	1.0
Win64 AMD64	mpich2-1.0-1-win64-amd64.zip	<a href="#">http</a>	<a href="#">ftp</a>	1.8MB	1.0
Win64 IA64	mpich2-1.0-1-win64-ia64.zip	<a href="#">http</a>	<a href="#">ftp</a>	3.3MB	1.0

In addition to the distributions provided by the MPICH2 Development Team, distributions of MPICH2 for other operating systems and packing environments are available below. Many thanks go out to those individuals who have graciously contributed their time and energy to create these distributions.

Platform	Author(s)	Email Address	Download		Size	Version
Debian	Zach Lowry	<a href="mailto:zach@zachlowry.net">zach@zachlowry.net</a>	mpich2_1.0-1_i386.deb	<a href="#">http</a>	<a href="#">ftp</a>	1.6MB 1.0

 [www.mpich.org](#)  Google 

## MPICH

High-Performance Portable MPI

Home About Downloads Documentation Support ABI Compatibility Initiative

**MPICH** is a high performance and widely portable implementation of the **Message Passing Interface (MPI)** standard.

MPICH and its derivatives, form the most widely used implementations of MPI in the world. They are used exclusively on nine of the top 10 supercomputers (November 2013 ranking), including the world's fastest supercomputer: Tianhe-2.

Download MPICH

### NEWS & EVENTS

**MPICH 3.1 released**

The MPICH team is pleased to announce the availability of a new stable release (mpich-3.1). This is a new ...

[Read More >>](#)

### LEARN ABOUT MPICH

[The documentation page](#) provides documents for installing MPICH, how to get started with MPI, and how to run MPI applications. It also includes tutorials, publications and other documents for developers.

[Read More >>](#)

### SUPPORT

[The support page](#) provides help for MPICH users and developers. There are links to frequently asked questions, support mailing lists and a trac system to report new bugs.

[Read More >>](#)

About Support News Documentation Downloads Publications Collaborators FAQ RSS Feed

# Mpich2 on Yarn

GitHub

This repository

Search or type a command

Explore

Features

Enterprise

Blog

Sign up

Sign in

PUBLIC



clarkyzl / mpich2-yarn

★ Star

24

Fork

14

Running MPICH2 on Yarn

79 commits

2 branches

0 releases

2 contributors



branch: master

mpich2-yarn / +

Merge pull request #30 from fredfsh/fredfsh-hadoop-2.4.0



clarkyzl authored on May 9

latest commit 0d703fa1b7



.settings

Make Eclipse work with JRE 1.6

2 years ago



mpich2-1.4.1p1

homedir: \$HOME -> /home/\$USER

15 days ago



src

DistinctContainersAllocator with AMRMClient

13 days ago



.classpath

Make Eclipse work with JRE 1.6

2 years ago



.project

issue #5 Use maven for compiling

2 years ago



README.md

Update README.md

11 months ago



pom.xml

ApplicationMasterProtocol -> AMRMClient

17 days ago



README.md

## mpich2-yarn

<> Code

Issues

9

Pull Requests

0

Wiki

Pulse

Graphs

Network

HTTPS clone URL

[https://github.com/](https://github.com/clarkyzl/mpich2-yarn)

You can clone with [HTTPS](#) or [Subversion](#).

Clone in Desktop

Download ZIP

amplab-extras.github.io/SparkR-pkg/

[View On GitHub](#)

DOWNLOADS

[TAR](#)

[ZIP](#)

## SparkR

R frontend for Spark

Project maintained by [amplab-extras](#)

Hosted on GitHub Pages — Theme by [mattgraham](#)

### R on Spark

SparkR is an R package that provides a light-weight frontend to use Apache Spark from R. SparkR exposes the Spark API through the `RDD` class and allows users to interactively run jobs from the R shell on a cluster.

### Features

#### RDDs as Distributed Lists

SparkR exposes the RDD API of Spark as distributed lists in R. For example we can read an input file from HDFS and process every line using `lapply` on a RDD.

```
sc <- sparkR.init("local")
lines <- textFile(sc, "hdfs://data.txt")
wordsPerLine <- lapply(lines, function(line) { length(unlist(strsplit(line, " ")) }) })
```

In addition to `lapply`, SparkR also allows closures to be applied on every partition using `lapplyWithPartition`. Other supported RDD functions include operations like `reduce`, `reduceByKey`, `groupByKey` and `collect`.

#### Serializing closures

SparkR automatically serializes the necessary variables to execute a function on the cluster. For example if you use some global variables in a function passed to `lapply`, SparkR will automatically capture these variables and copy them to the cluster. An example of using a random weight vector to initialize a matrix is shown below



## Mahout News

---

### 25 April 2014 - Goodbye MapReduce

The Mahout community decided to move its codebase onto modern data processing systems that offer a richer programming model and more efficient execution than Hadoop MapReduce. **Mahout will therefore reject new MapReduce algorithm implementations from now on.** We will however keep our widely used MapReduce algorithms in the codebase and maintain them.

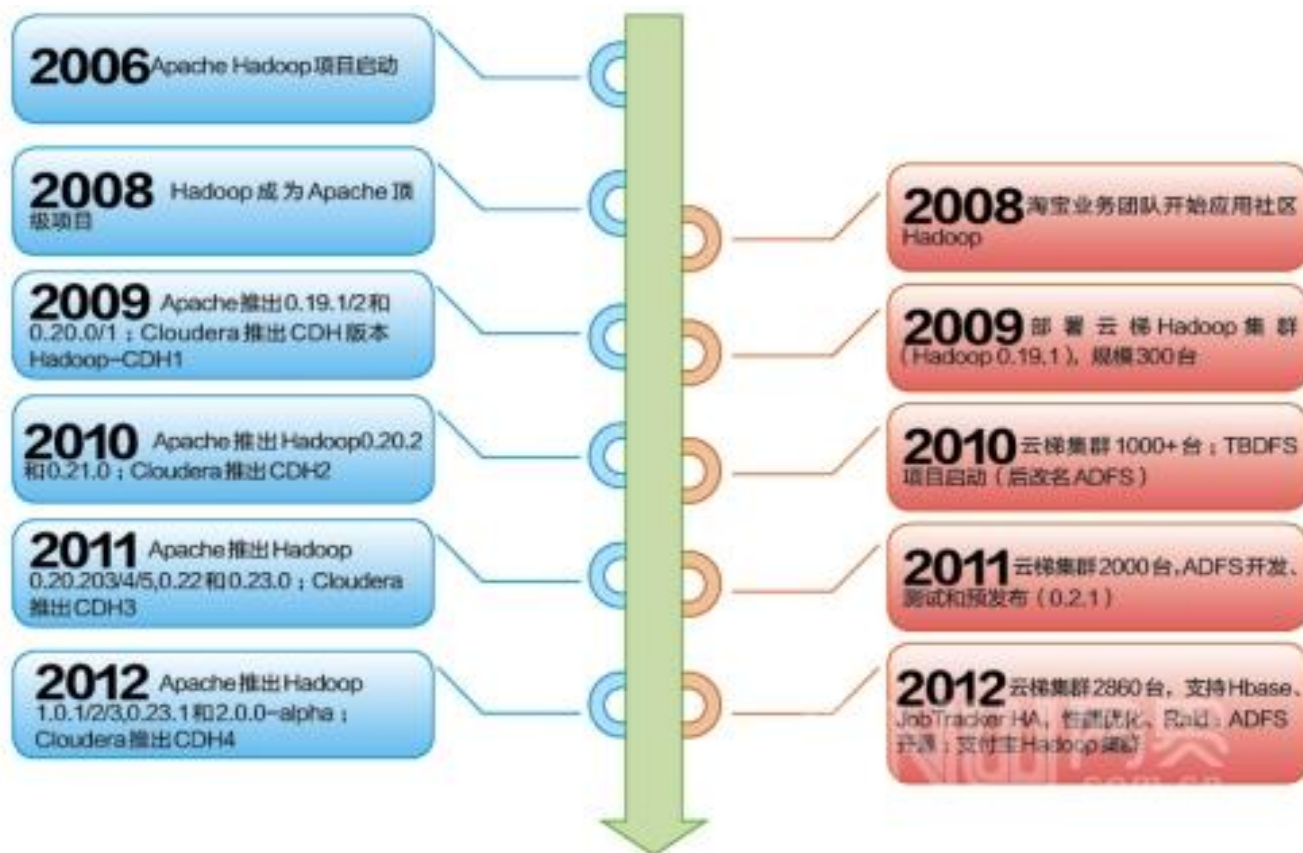
We are building our future implementations on top of a [DSL for linear algebraic operations](#) which has been developed over the last months. Programs written in this DSL are automatically optimized and executed in parallel on [Apache Spark](#).

Furthermore, there is an experimental contribution undergoing which aims to integrate the [h2o platform](#) into Mahout.

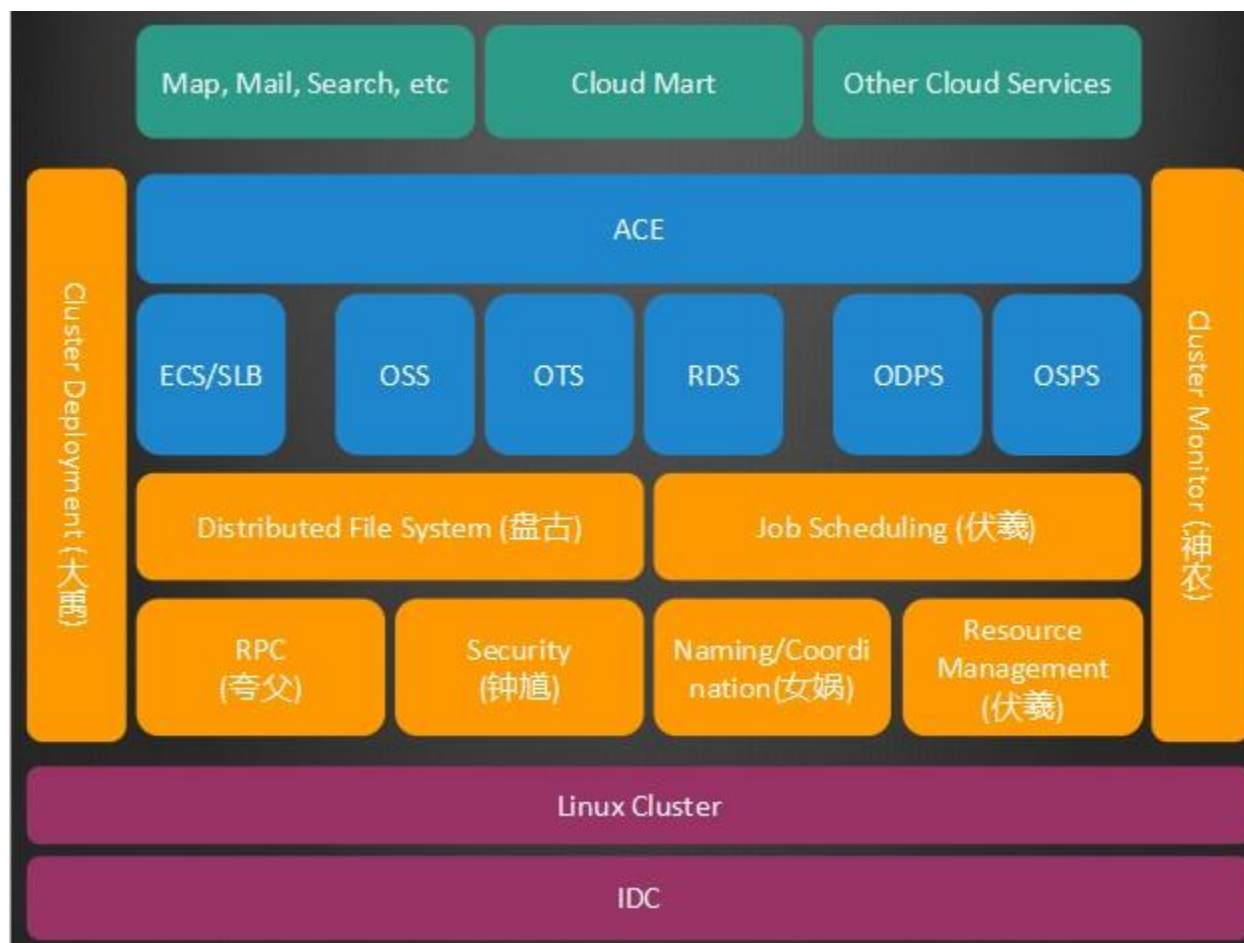
### 1 February 2014 - Apache Mahout 0.9 released

Apache Mahout has reached version 0.9. All developers are encouraged to begin using version 0.9. Highlights include:

- New and improved Mahout website based on Apache CMS - [MAHOUT-1245](#)
- Early implementation of a Multi Layer Perceptron (MLP) classifier - [MAHOUT-1265](#)
- Scala DSL Bindings for Mahout Math Linear Algebra. See this [blogpost](#) and [MAHOUT-1297](#)
- Recommenders as Search. See [<https://github.com/pferrel/solr-recommender>] and [MAHOUT-1288](#)
- Support for easy functional Matrix views and derivatives - [MAHOUT-1300](#)
- JSON output format for ClusterDumper - [MAHOUT-1343](#)
- Enabled randomised testing for all Mahout modules using Carrot RandomizedRunner - [MAHOUT-1345](#)
- Online Algorithm for computing accurate Quantiles using 1-dimensional Clustering - See this [pdf](#) and [MAHOUT-1361](#)



- <http://www.csdn.net/article/2013-12-05/2817724-bdtdc2013-aliyun>



# 放弃云梯的原因

- 阿里集团高层是决心自主研发一套有 自主知识产权的分布式计算系统，即后来代号为“飞天”的大规模分布式计算系统。而基于开源Hadoop技术的集群，最初只是被定位 成一个临时的、有过渡性质的系统，目的是让淘宝的业务人员提前熟悉和使用分布式计算系统，待“飞天”成熟后再将业务移植过来。这个系统是为“飞天”铺路的，所以叫“云梯”，隐含奉献的意思。
- 就来Hadoop的蓬勃发展，演变为云梯与飞天并行存在的局面，但最终要作取舍。
- 深受现有Hadoop版本之苦，由于不是Hadoop项目管理委员会的成员，Hadoop开源社区的发展并不受阿里的控制和影响，这使得阿里不能很好地定制Hadoop，在研发上受制颇多。
- 阿里自主研发的飞天分布式平台会成为阿里数据平台的主力，作为一家有野心的互联网公司，阿里巴巴做出这样的技术路线选择虽在意料之外，却也在情理之中。

- Splunk
- Palantir
- Datasift
- Decide.com

- Splunk软件平台可以实时对任何APP、服务器或网络设备的机器数据进行索引、监控与分析，并将结果生成图形化报表，在此基础上帮助客户避免服务性能降低或中断。这些机器数据可以是日志、配置文件、消息和告警等，既可以来自本地也可以来自云，并且是动辄TB级别的、部署于成万千上万台服务器的数据。
- Splunk的业务迎合了大数据时代企业对数据应用的需求,其业务功能主要分为五大块：IT运营、应用管理、安全合规、网络智能与商业分析。此外，Splunk的搜索功能异常强大，被称为“**Google for IT**”。
- Splunk的产品有免费和收费版，二者最主要的差别在于每天的索引容量大小(索引是搜索功能的基础)，免费版每天最大为500M。如果需要海量索引量及更多的功能，例如分散式搜寻(Distributed Search)、排程告警(Schedule Alert)、权限(Access Control)等功能，则需要购买企业版，不同索引量的价格不同。

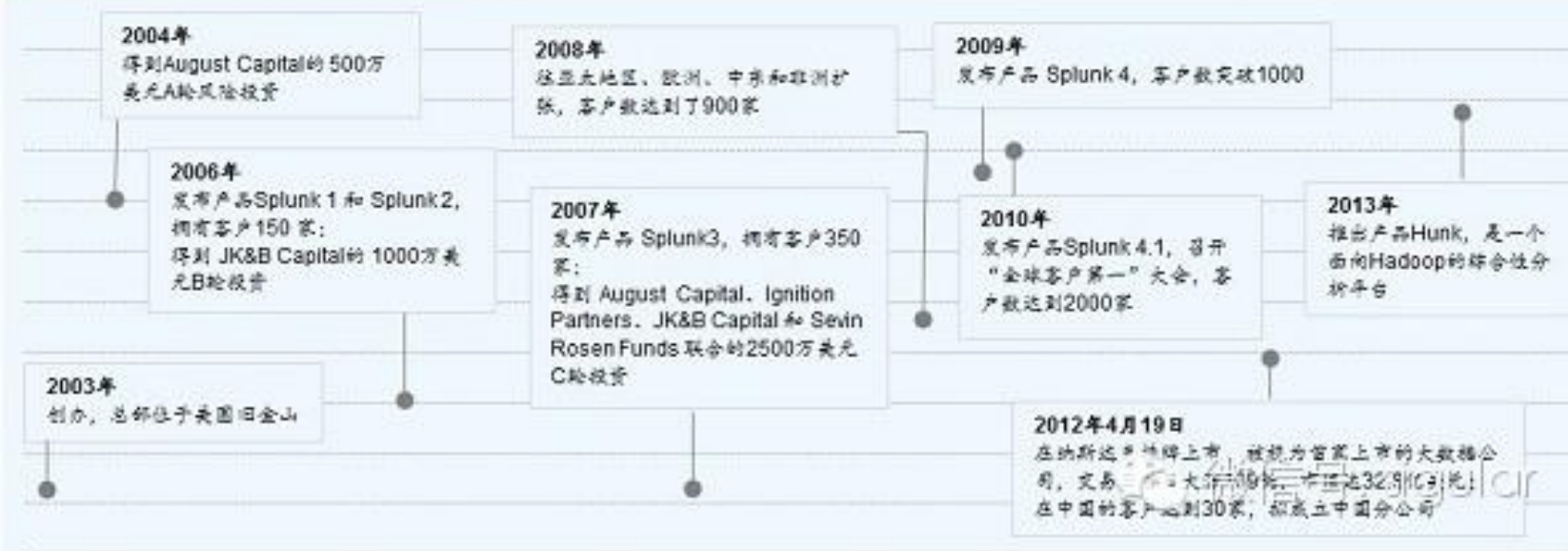
- 截至2014年1月，有7000多个客户在使用Splunk的产品和服务，其中70个客户是财富100企业，8个客户是全球10大公司。客户所在的行业从传统行业、科技行业到在线服务行业无所不有，下图所举例子显示其客户所覆盖的行业。在中国市场，Splunk的业务主要集中在电信、保险和银行业等，例如银联支付、民生保险、百联支付、国美电器、中国移动和中国电信等。

教育行业	金融行业	零售行业	传媒行业
哈佛大学、纽约大学	美国银行、JP 摩根	Freshdirect、梅西百货	BBC、HBO 电视网
制造行业	科技行业	在线服务行业	微信号: datagular
通用电力公司、实耐宝	思科、摩托罗拉	Expedia、Paypal	美国凤凰城、USAID





## Splunk里程碑

2008年以后，Splunk进入扩张期，客户数在2009年突破1000，2010年突破2000，2011年突破3000，在2012年上市前达到3700，其中来自中国的有30余家。



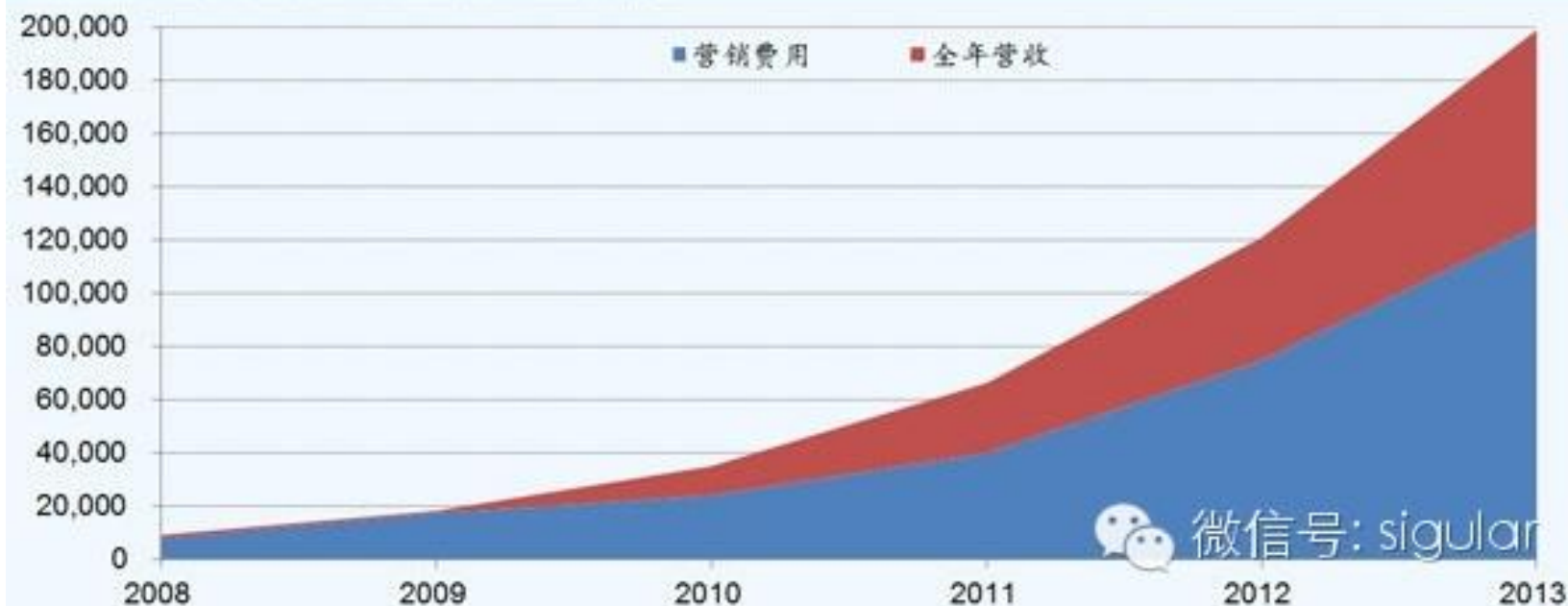


案例二：Chandler Police Department	
	Chandler Police Department, 美国亚利桑那州的城市 Chandler 的警察部门, 有 320 个官员和 150 个雇员, 为该市 25 万居民服务。
问题	该部门需要对与警务相关的加密记录、数据和程序进行日常监控, 这些数据主要来自其 LDAP (轻量级目录访问协议) 和网络服务器, 以及 RMS (记录管理系统) 生成的机器日志, 这些系统的功能都有待优化和整合。
方案	Splunk 可以优化部门的原有数据, 如将 RMS 里的数据按给定的时间表进行索引、加快数据提取的速度等; 分析数据, 如从累积的数据中分析该市的警务反应是否合格以进一步提升服务、分析案件多发地带以形成更有效的预警机制等。Splunk 提供的服务不仅可以节约工作人员的大量时间与精力, 还可以提供更快速更有效的反应机制, 从整体上提升该部门的服务质量, 进而提升该市的治安环境。

案例三：UniCredit Business Integrated Solutions	
	UniCredit Business Integrated Solutions (联合信贷商务集成解决方案), 是意大利联合信贷集团推出的可使用 NFC 手机 (近距离通信手机) 进行支付的解决方案, 该方案是为多应用手机钱包, 所有的服务集成于一个虚拟钱包中。
问题	对于金融服务行业来说, 如何向客户保证其隐私和安全是为关键。这一多功能手机钱包在不同操作平台上运营着大约 1000 个应用, 显然, 其面临的主要问题是搜集、存储与分析每天持续生成的海量机器数据, 并鉴定其中的安全隐患、检测其中的欺诈交易。
方案	Splunk 极易使用, 无需任何特殊的 IT 技能。该方案的每一个部门都接入 Splunk 后, 少量数据的整合与监控也变得更简单更快。通过实时监控和事件排查等功能, Splunk 减少了该方案的 MTTI (平均调查时间)、MTTR (平均解决时间) 以及其它 IT 性能问题, 40% 的事件可以在用户发现之前得到排查, 为问题排查节省的时间至少降低了 70%。公司已决定将每天的索引量从 250G 升级到 2.8TB。

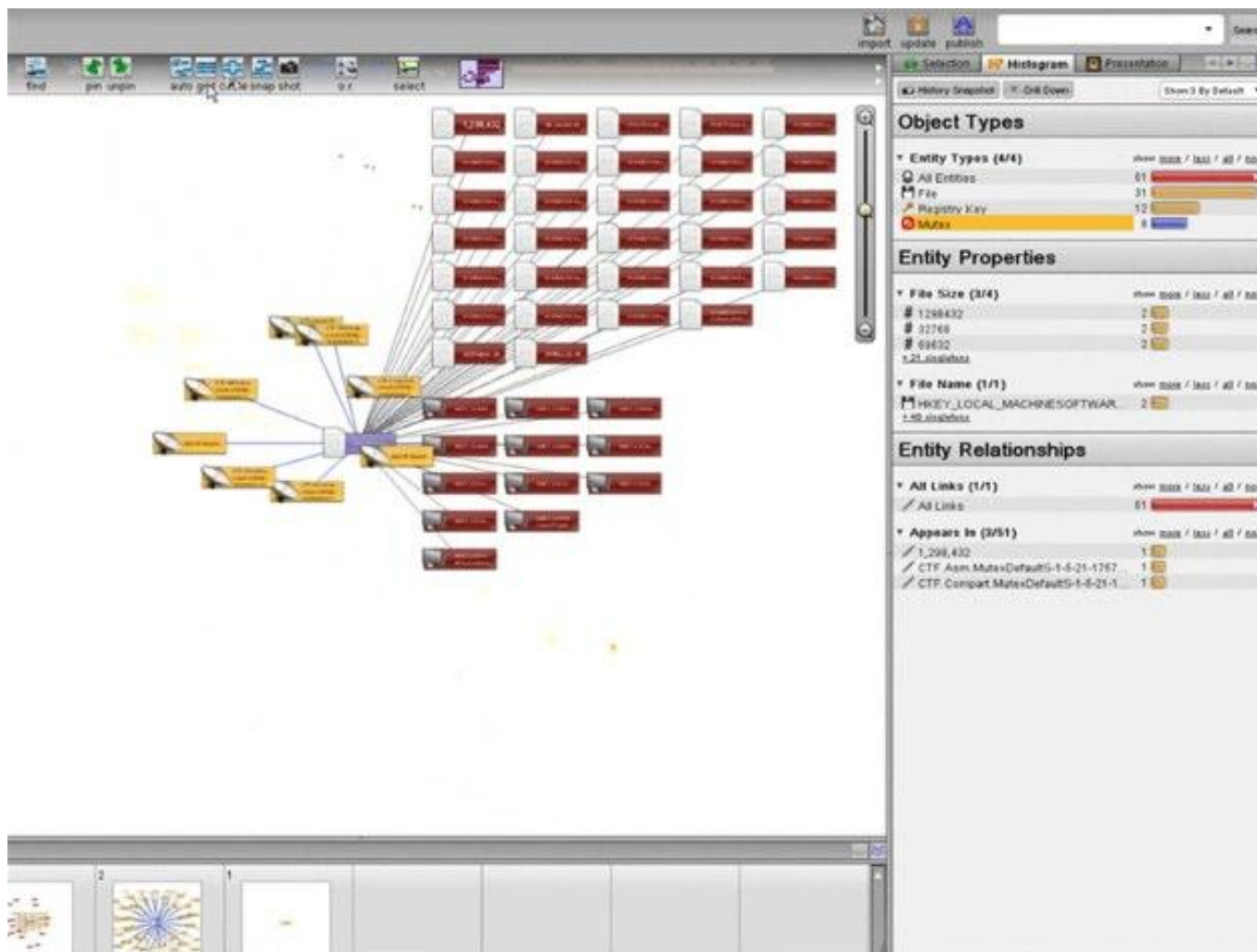
## Splunk全年营收与营销支出(2008-2013)

近年全年营收与营销支出均呈现明显的上升趋势。

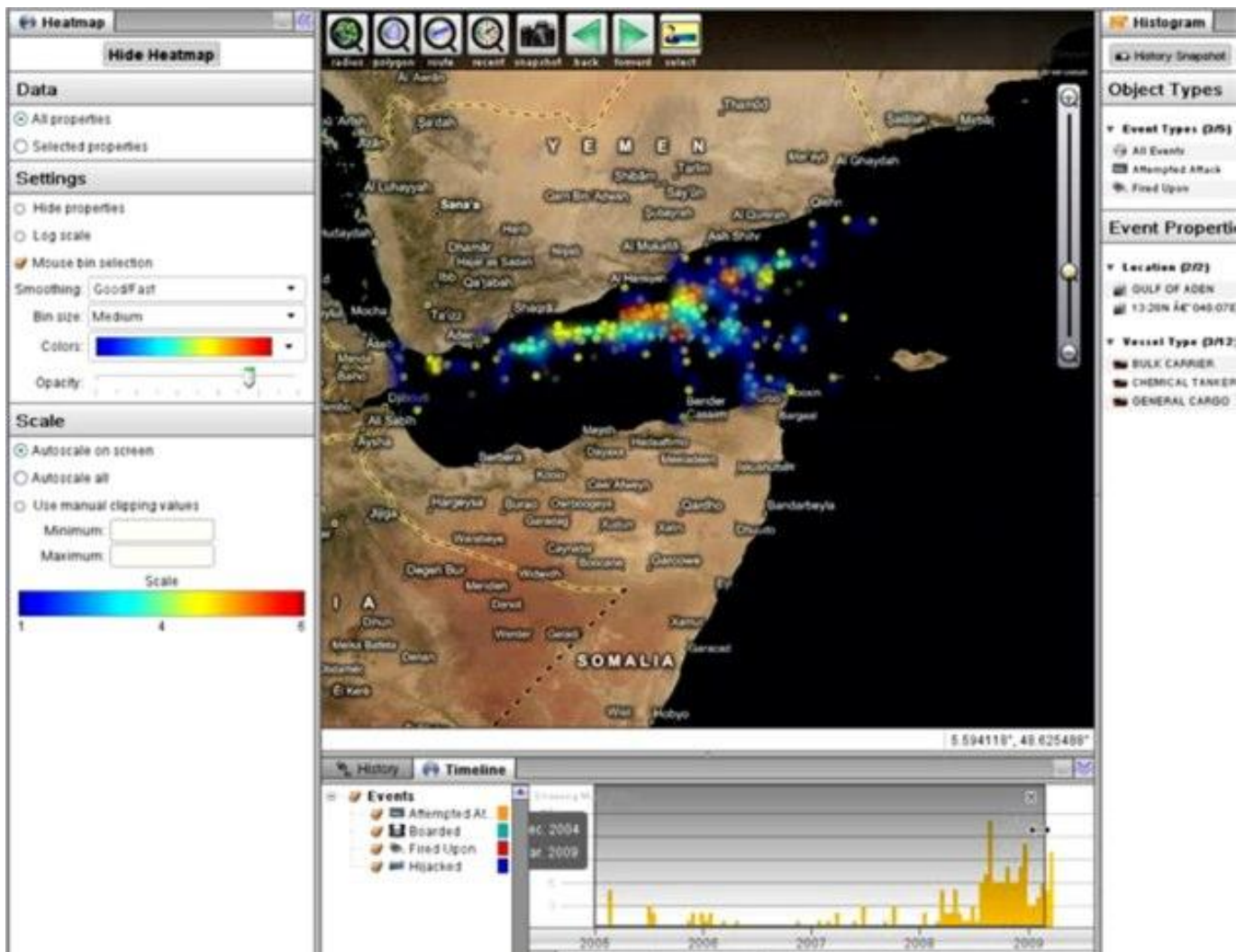


- 2004年成立，目前市值超过90亿美元
- 他们提供一个很难解释清的产品，帮助用户用可视化的方式在高度复杂的大型数据集 中进行探索，从而直观地发现内里的规律和未知因素。
- 主要用户都在首都华盛顿，来自政府的业务占到了 70%，其余业务主要来自私人金融 机构。用于反恐战争、网络间谍、经济刺激计划执行情况（实际上 Palantir 帮助政府 发现刺激计划中的诈骗行为）、医保、乃至自然灾害等方面派上了大用场
- Palantir 帮助多伦多大学 Munk 全球事务学院的科研人员们发现了一个名为“影子网 络（Shadow Network）”的网络间谍组织，该组织当时正在从印度国防部窃取机密 资料。可以这样说，哪里有危机，哪里就有 Palantir 的用武之地。

# 视眼石项目截图



# 视眼石画出的索马里海盗攻击热点





- 各种视频演示：<http://www.palantir.com/library/>
- 旗舰产品精益求精的执着精神在 Palantir 处处可见。比如，公司每个月发布一次软件更新（政府类产品），工程师们利用元素周期表中的元素来命名这些更新，并设计专门的 T 恤加以纪念。以工程师为本的公司文化进一步强化了这种精神，实际上所有员工都是 27 岁上下的工程师。
- 没有公关，没有销售，没有营销，创始人Karp 坚称这些永远都不会有。他说自己非常喜欢让口碑推动业务、媒体和销售。

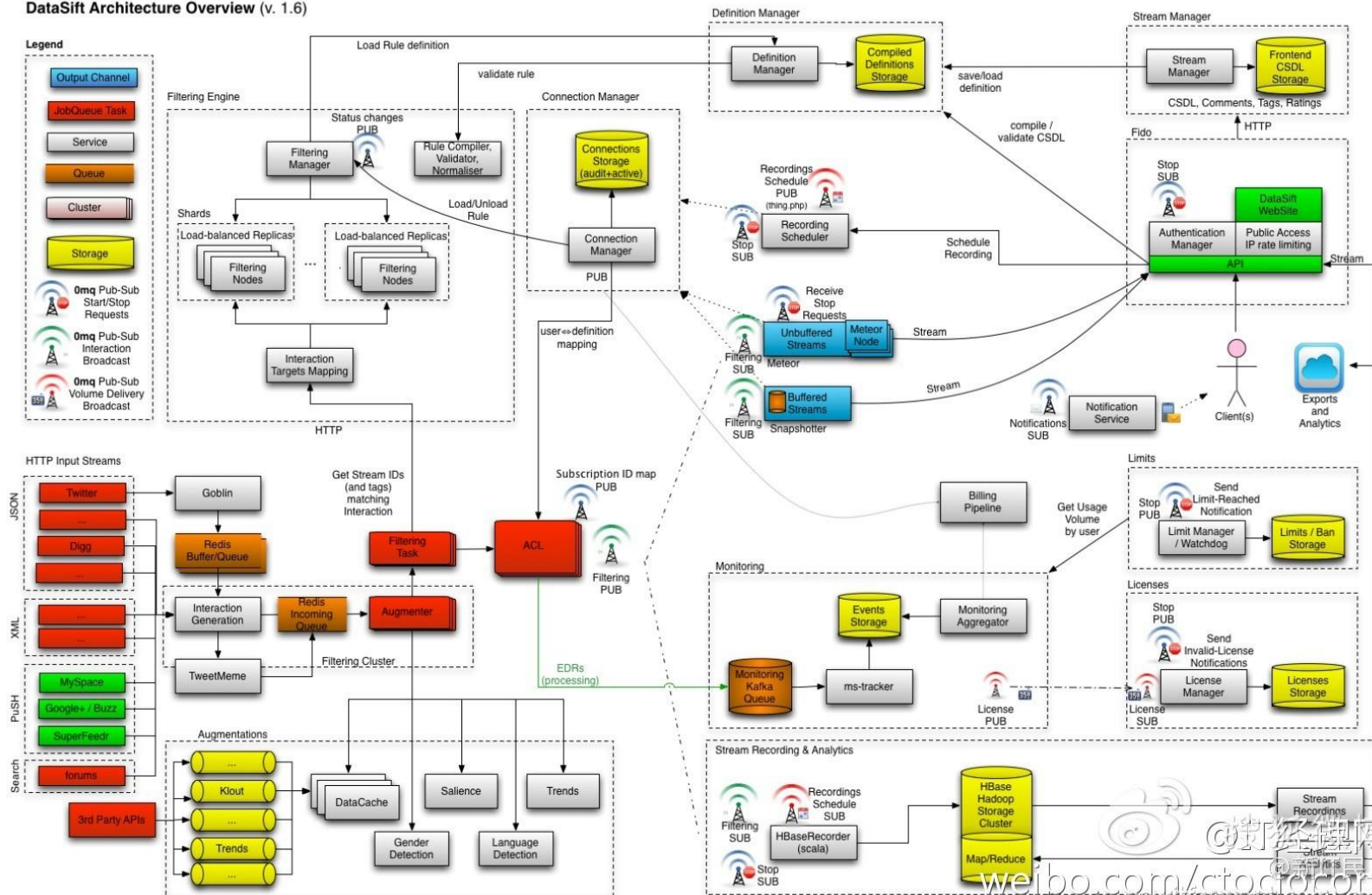
- DataSift前身是Twitter信息过滤平台Tweetmeme，初期的业务是帮助开发者和第三方获取并分析来自 Twitter、Facebook、Tumblr 等社交网络的数据。如今，除了社交网络，DataSift 更深入社交媒体领域。
- DataSift的定位是实时社交数据挖掘平台，在互联网上向大规模用户提供Twitter数据分析服务。
- Twitter开放其数据管道Firehose对于社交大数据分析来说无疑是一个晴天大利好。利用Twitter实时数据你几乎能进行各种数据分析，从奥斯卡电影人气到美国总统支持率，再到产品用户满意度分析，可谓一座不设防的数据大金矿。但是掘金Twitter “快数据” 也对分析系统提出了很高要求，DataSift是少数能吃下Twitter数据的顶级社会化分析机器之一，DataSift从Twitter购买了多年的数据同步授权，能够访问所有Twitter管道数据，并将子集卖给第三方，主要是企业客户。目前只有Gnip获得了同样的授权。

- 运行于SSD固态硬盘之上的MySQL ( Percona server )
- HBase集群 ( 目前约30个Hadoop节点, 400TB存储 )
- Memcached ( cache )
- Redis ( 依然用于一些内部队列、但也许很快将弃用 )
- 每秒实时挖掘12万条Twitter内容



# Datasift架构

DataSift Architecture Overview (v. 1.6)



- 2013年12月4日，苹果以2亿美元收购了社交媒体分析公司Topsy。据Techcrunch (TC) 报道，社交媒体分析公司DataSift宣布完成4200万美元的C轮融资。
- 目前该公司已在全球40多个国家拥有超过1000家企业客户，包括彭博社，道琼斯，CBS Interactive，Dell、Marketwired、Dachis Group、Conversocial、SecondSync、HootSuite 和 Simply Measured等。



## Samsung PN 51" Plasma

PN51D6500 • 1080p

Model » released Mar 2011 • 3 months old

Reviews » ★★★★★: 3

Prices » from 14 stores • \$1,348.39 (6% off) at Amazon

**\$1,348.39**

**Set Alert**



### Wait for prices to drop \$92

Prediction: Prices will drop or hold steady (\$92 on avg.) **87% Confidence**



- Decide.com 的目的是解决两个问题：电子产品更新速度过快，市场价格变化过快。通过解决这两个问题，他们希望给潜在购物者提供最好的购物时机建议。
- “线性分析模式”，将成千上万个电子产品加入到了自己构建的数据库。并且去爬许许多多的技术博客和网站以获取产品发布消息和传闻，最终运用“**先进的机器学习以及语意挖掘算法**”来预测未来的产品发布时间。
- 另外除了对产品发布时间进行预测，他们还会利用专有价格预测算法通过综合考虑上亿条价格波动信息和超过 4 0 个不同的价格影响因素来对价格做出全面的价格预测。

- 用户使用该网站则非常简单，登入 Decide.com，在搜索框中输入具体想购买的电子产品，然后 Decide.com 便会根据自己专有技术预测给你返回是应当购买还是应当等一等的建议。如果用户想要购买，则可继续点击进入他们选中的卖家进行购买。如果根据 Decide.com 给出的建议，用户暂时不想购买，那么他们则可以观看价格时间线，并给自己设定一个价格下降的通知以不错过自己理想的购买价格。
- 其整个产品发布和价格预测体系都搭建在对海量数据的挖掘和分析处理上。而为了证明自己的预测是有效的（至少是局部有效的），从今天开始他们每天会筛选 10 项产品进行保价赔付，也就是说用户购买这 10 项产品两周之类出现跌价，他们会进行差额赔付。
- 不过，我们无法得知这 10 项产品到底是人工操作的还是真的是机器推荐的。如果真的源于 Decide.com 对数据的分析作出的推荐，那么其价格预测体系应该算是初有成效了。另外值得指出的是 Decide.com 的盈利模式是收取零售商的佣金，比如其收取 Amazon 的佣金比例就高达 4%。那么这其中是否会有人工操作呢？

- [http://en.wikipedia.org/wiki/Oren\\_Etzioni](http://en.wikipedia.org/wiki/Oren_Etzioni)
- Oren Etzioni is an American entrepreneur and professor of Computer Science and Executive Director of the Allen Institute for Artificial Intelligence. He joined the University of Washington faculty in 1991, where he became the Washington Research Foundation Entrepreneurship Professor in the Department of Computer Science and Engineering. In May 2005, he founded and became the director of the University's Turing Center. The Center investigates problems in data mining, natural language processing, the Semantic Web and other web search topics. He coined the term machine reading and he created the first commercial comparison shopping agent.

- Etzioni is an entrepreneur who has founded or co-founded several business ventures, including **MetaCrawler** (bought by Infospace), **Netbot** (bought by Excite), and **ClearForest** (bought by Reuters). He founded **Farecast**, a travel metasearch and price prediction site, which was acquired by Microsoft in 2008. He co-founded **Decide**, a company whose website Decide.com helped consumers make buying decisions using previous price history and recommendations from other users. **Decide.com was bought by eBay** in September, 2013. He is also a venture partner at the Madrona Venture Group





# 被ebay收购





- **Dataguru（炼数成金）是专业数据分析网站，提供教育，媒体，内容，社区，出版，数据分析业务等服务。我们的课程采用新兴的互联网教育形式，独创地发展了逆向收费式网络培训课程模式。既继承传统教育重学习氛围，重竞争压力的特点，同时又发挥互联网的威力打破时空限制，把天南地北志同道合的朋友组织在一起交流学习，使到原先孤立的学习个体组合成有组织的探索力量。并且把原先动辄成千上万的学习成本，直线下降至百元范围，造福大众。我们的目标是：低成本传播高价值知识，构架中国第一的网上知识流转阵地。**
- **关于逆向收费式网络的详情，请看我们的培训网站 <http://edu.dataguru.cn>**



# Thanks

## FAQ时间