基于实时计算的Web 分析系统的研究与开发

# 研究背景及意义

随着互联网的迅速发展，运营商面临的竞争日益激烈，挖掘用户行为产生的数据，可帮助运营商监控网站运行状态，更好的改善网站结构，做更好的业务和营销决策。现在很多公司每天都会产生数以TB级的大数据，如何对这些数据进行挖掘，分析成了很重要的课题。大数据时代，传统的web分析系统架构难以适应海量数据挖掘的要求。同时实时监控网站运行情况的需求也日益剧增。例如电子商务网站，需要处理并且挖掘用户行为产生的数据，产生推荐，从而带来更多的流量和收益，而每天处理海量的用户数据，需要一个低延时高可靠的实时流式分布式计算系统。再如社交网站，大家每天都会去社交网站是为了看看现在发生了什么，周围人在做什么，流式计算可以把用户关注的热点聚合，实时反馈给用户，从而达到一个圈子的聚合效果。

本课题研究工作是基于对公司网站，实现用户行为数据的采集，数据存储，以及实时分析。

# 与本课题相关的研究现状

## Web分析发展过程

美国网站分析协会WAA(Web Analytics Association)（现已改名为DAA:Digital Analytics Association）对网站分析的官方定义

“Web Analytics is the measurement, collection, analysis and reporting of Internet data for the purposes of understanding and optimizing Web usage.”

网站分析工具的鼻祖是[WebTrends](http://webtrends.com/)在1993年美国波特兰成立，基于日志分析用户行为。在1996年第一个网站计数器[Web-Counter](http://www.hitwebcounter.com/)诞生，之后在各大网站都开始出现访客计数器。 2005年Google收购Urchin之后推出[Google Analytics](http://www.google.com/analytics/)，2008年10月Yahoo!收购IndexTools后推出[Yahoo! Web Analytics](http://web.analytics.yahoo.com/" \t "_blank)，2011年Google宣布推出针对大客户的付费版Google Analytics Premium。同时国内的百度公司也推出了百度统计，并在2011年将百度统计开放给任何人使用，全部免费。

Web分析的关键指标包括：网站基础指标、流量质量指标、转化指标。

## Web分析流程

最早提出Web挖掘的是Oren Etzioni [2]，在他1996年的论文中声称，Web挖掘是利用数据挖掘技术从互联网的文档和服务中来自动发现可用信息。

--An Approach for Frequent Access Pattern Identification in Web Usage Mining

就挖掘的内容而言，web挖掘Web 挖掘可分为 Web 内容挖掘、Web 结构挖掘和 Web 使用记录挖掘。

--- Formation of Algorithms Module and Dynamic Interface in Web Mining and Storage Process to Retrieve Metadata

Web使用记录挖掘，也就是本课题研究的Web分析，流程可分为源数据收集、数据存储、数据挖掘和模式分析 3 个阶段。

### 数据收集

当访客和web站点进行交互时,主要用来收集数据的方法有:web日志[6]、Javascrit标记法。而有些公司的高级电子商务软件内置了数据收集机制,如收集重要的业务事件和背景数据的事件日志。有时候需要数据收集的方法不止一种,可能会多种方法同时使用。

1. Web 日志技术

web日志[61从web出现的时候起就一直是数据收集的最原始数据源,它最初是被开发出用来捕获Web服务器产生的错误。随着分析的需求从基于技术转向基于营销,它的功能慢慢地被“强化”为捕获更多的数据。当访客客户端同浏览器建立连接后，浏览器通过 socket向网站服务器发送请求，服务器接收到请求，然后根据请求找到客户端需要的文件，然后在日志中记录下这个访客的请求信息，最后服务器将浏览器请求的文件发送给浏览器。

典型的网站日志文件，通常包含如下信息：

* 远程主机的 IP 地址或名字；
* 访客登录名；
* 访客登录全名；
* 请求所发生的日期时间；
* 请求方法；
* 请求的文件地址；
* 请求所遵守的协议；
* 请求的状态；
* 被请求文档的大小。

服务器日志分析法的优点在于他分析的数据与终端无关，即无论访问网站的是手机，电脑还是其他终端，他都能够真是地记录访客信息；服务器日志分析的数据不依赖于第三方的支持，只要有访客访问服务器，日志就会被记录；服务器日志分析的数据源是网站的日志，所以他的数据源是最容易获得的；采用服务器日志分析法分析网站的结果数据是可恢复的，只要有原始日志，即使分析结果丢失，也可以重新恢复；由于服务器日志是记录服务器的响应行为，所以她还能记录搜索引擎机器人在服务器上访问行为；还有一点就是用服务器日志法分析网站不怕客户端防火墙，由于数据源是从服务器上直接获得的，与访问终端无关，所以客户端防火墙的设置不影响数据收集。然而，日志分析法也存在缺点，只有访客访问服务器才会产生日志，但是随着代理服务器缓存和浏览器缓存的诞生，本来访客可能访问服务器的情况由于缓存的原因而未发生，导致日志记录的访问量少于实际访问量。日志分析法不能精确地辨别独立访客，由于日志分析法主要是根据访客 IP 地址来区分访客的，但是这样做导致同一个局域网下的所有访客被人为是一个访客。显然这种分析方法相对来说准确度很低。

1. JavaScript标记技术

JavaScript标记[l一2]目前是业界最普遍和最受欢迎的技术,大部分的供应商和web分析的解决方案都依靠JavascriPt标记来收集数据。JavaScript标记法网站分析是在网页源文件中插入一小段可执行的 JavaScript程序代码当网页被下载到用户客户端的浏览器上时，这段 JavaScript代码就会被执行。随后，它就会如实的把访客在页面上的访问行为不间断的发送给这个网站分析所对应的服务器.

JavaScript标记法网站分析是最流行的一种分析技术。同日志分析法相比，页面标记法不怕代理服务器缓存和浏览器缓存的影响，由于页面标记的源代码是放在网页源文件中的，即使网页被代理服务器的缓存或客户端浏览器的缓存保存下来，页面标记的源代码也会跟着一起被保存，并且当浏览器载入页面的时一起被执行；页面标记法主要依赖于Cookie 中的记录来辨识访客的身份信息，同仅依赖于访客 IP 地址辨识访客身份的日志分析法相比，页面标记法具有相对准确的访客记录；页面标记法网站分析可以只对网站的某个具体的页面进行分析，灵活性更大；它具有更好的实时性，基本上在访客访问网站的同时就完成了页面统计，

### 数据存储

数据库存储技术已经走过三十几年的发展，各种数据库系统的设计为满足不同尺度的数据规模以及复杂变化的应用需求。随着近年来电子商务网站、超大规模的门户网站及高并发性SNS纯WEB2.0动态网站的兴起，越来越多的系统对数据库提出了高并发读写、高效率存储和访问、高扩展和高可用性等需求，而传统的关系型数据库无法满足这些需求。在 WEB2.0的网站中，关系数据库大部分都出现了瓶颈。在磁盘IO、数据库可扩展上都花费了开发人员相当多的精力来优化，比如做分表分库、主从复制、异构复制等等[1]。而非关系型数据库的一个共同特点都是去掉关系数据库的关系型特性。数据之间无关系，这样就非常容易扩展。也无形之间，在架构的层面上带来了可扩展的能力。接下来针对其三种常见类型，键值存储、列存储、以及文档存储。

1. 键值存储

即Key-Value存储，简称**KV存储**。它的数据按照键值对的形式进行组织，索引和存储。每个key都是唯一的，客户端根据key来获取value值。键值数据库的出现，主要是受到了Amazon's Dynamo的影响。在Dynamo中，数据被分区存到多个服务器中并且有多个备份。可扩展性和持久性依赖于两个关键机制：分区与复制、数据版本控制。

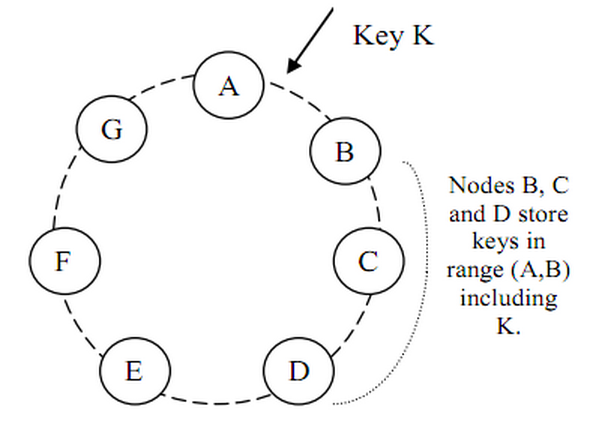


图 1 **Dynamo 环上的分区和复制**

* 分区与复制：

Dynamo对所有保存的数据Key做MD5哈希生成128位的特征码，将所有128位的数值首尾相接就形成了图1的环（Dynamo Ring），Dynamo Ring被节点（Node）切割形成了若干数据段（Key Range）。根据特征码将保存于Dynamo的数据对象分布到不同的Key-Range上，集群中的服务器（Host）各持有一个Node，负责存储当前Node与上一个Node划分出的Key-Range上的数据对象。伴随Dynamo存储的数据不断膨胀，必然需要添加更多的Node来把Dynamo Ring切分成更小范围以扩张系统的性能。添加的Node会从其它相邻的节点上“偷取”一部分Key Range。这样在添加Node时受影响的就只有与之直接毗邻的那个Node了。比如我们在A、B之间添加Node A’，则只有Node 会受到影响将自己的一部分数据转移给A’。    同样的在一个Node发生故障或者移除Node的时候，受影响的也只有它直接毗邻的那个Node。

------From：http://my.oschina.net/pangyangyang/blog/151093

* 数据版本控制：

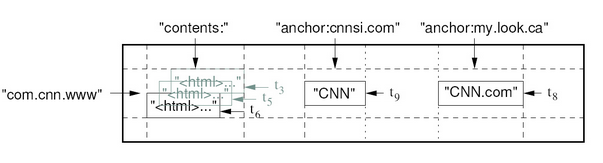
在Dynamo中，由于并行的更新和错误的发生，可以导致某个对象的版本有分枝出现。此时系统无法协调这个对象的版本，必须保同的版本，等待客户端来决定最终的版本。Dynamo通过这种方式来保证尽量不丢失任何的更新。Dynamo中采用向量时钟来解决这个问题。每个对象都有一个向量列表，列表中每个元素是一个（node, counter）对，表示该对象包含了node上第counter次更新的结果。由于正常情况下，更新都是在preference list中的前N个节点上发生，所以向量列表的长度不会太长。万一出现一些异常情况，可能会导致向量列表的增长，Dynamo采用了一种简单的方式来处理这种问题，当向量列表长度超过某个定值的时候（例如10），把最早的向量（node, counter）对删除，为此，每个向量还需要带有一个时间戮，表示该更新发生的时间。删除向量可能会导致某些更新丢失，但一般来说，Dynamo在生产环境中很少遇到这样的问题。

其他的键值数据库包括：Voldemort [138],Redis [139], Tokyo Cabinet [140] , Tokyo Tyrant [141],Memcached [142] , MemcacheDB [143], Riak [144],和 Scalaris [145]. Voldemort, Riak, tokyo Cabinet and Memecached 可以存储在内存中也可以存储在磁盘上。其他的存储在内存中，并提供磁盘恢复，或者依赖复制来恢复数据。

1. 列**式**存储

**列式[数据库](http://baike.baidu.com/view/1088.htm" \t "_blank)**是以列相关[存储](http://baike.baidu.com/view/87682.htm" \t "_blank)架构进行数据存储的[数据库](http://baike.baidu.com/view/1088.htm)，为了实现可扩展性，行与列被分割存放到个多个节点中。

BigTable是一个分布式存储系统，它可以支持扩展到很大尺寸的数据：PB级别的数据，包含几千个商业服务器。Google的许多项目都存储在BigTable中，包括WEB索引、Google Earth 和Google Finance。BigTable是一个稀疏的、分布的、永久的多维排序图。采用行键盘（row key）、列键（column key）和时间戳（timestamp）对图进行索引。图中的每个值都是未经解释的字节数组。

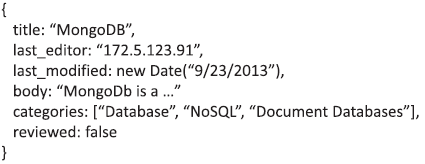


其他的列**式**存储数据库包括Facebook的Cassandra，HBase，以及Hyper-table等。列**式存储都借鉴自Google的Bigtable，每一行数据的各项被存储在不同的列中，而每一列中每一个数据都包含时间戳属性，这样列中的同一个数据项的多个版本都能保存下来。**

**--- from 书**

1. 文档存储

相对于键值数据库，文档数据库支持更复杂的数据结构。文档数据库可以看作是键值数据库的升级版，允许在存储的值中再嵌套键值。文档数据库主要包括MongoDB,CouchDB,SimpleDB,他们都是以JSON来存储数据，最主要的区别是数据复制以及一致性机制。



* 数据复制与切分

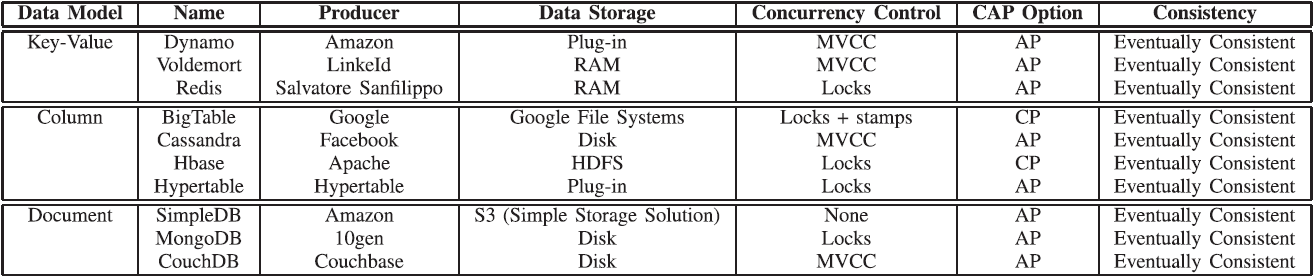
mongodb的复制至少需要两个实例。其中一个是主节点，负责处理客户端请求，其余的都是从节点，负责复制主节点上的数据。主节点记录在其上的所有操作oplog，从节点定期轮询主节点获取这些操作，然后对自己的数据副本执行这些操作，从而保证从节点的数据与主节点一致。SimpleDB复制数据到不同数据中心的不同机器上以保证数据安全并提升性能。CouchDB当前没有分片机制，使用乐观的复制实现可扩展性，每个CouchDB的数据库可以同步到另一个实例;因此，可以建立任何类型的复制拓扑。

* 一致性

MongoDB和SimpleDB没有并发版本控制和事务管理机制，但他们提供最终一致性。CouchDB的一致性根据配置而不同，对于主-主，CouchDB提供最终一致性，对于主-从，CouchDB支持强一致性。

1. 各种类型数据库对比

数据模型有着各自的优缺点，它们适用于不同的领域。不管是选择关系模型，还是非关系模型，都要根据实际应用的场景做出选择。也许你会发现单一的数据模型不能满足你的解决方案，许多大型应用可能需要集成多种数据模型



### 数据处理

Batch定时任务，到hadoop,到实时计算。

Generally the web mining tasks can be classified into threecategories: web content mining, web structure mining and webusage mining. In addition, there are two other differentapproaches to categorize web mining. In both, the abovecategories are reduced from three to two: web content miningand web usage mining. In one, web structure is treated as partof web Content while in the other; web usage is treated as partof web Structure.

--- Formation of Algorithms Module and Dynamic Interface in Web Mining and Storage Process to Retrieve Metadata

Web usage mining includes three main steps: Data

Preprocessing, Knowledge Extraction and analysis of

extracted results.

Data Log fetch

Data clean

Pageview Identification

Computing the Reference Length

User Identification

Session Identification

**A NewClustering and Preprocessing for Web Log Mining**

Semantic Data Mining

Web Usage Mining A Review on Process, Methods and Techniques --- 匹配度很高

RTIC-C: A Big Data System for Massive Traffic Information Mining – 较高

Toward Scalable Systems for Big Data Analytics:

A Technology Tutorial 最牛逼的文章

# 研究现状

## user traversals

## Frequent Pattern Tree Algorithms

# 研究目标

# 关键技术

数据不丢失

笔记：

Spark RDD可容错