atitit .大数据之道.v2 rb2.docx

paip.论大数据的方法,技术.attilax总结.

[1. 大数据是什么 4v 大规模 非结构化 2](#_Toc14413)

[2. 大数据技术综述 巨量数据中 快速获得有价值信息的技术 3](#_Toc25169)

[3. 大数据的格式种类:普通数据 以及图片视频半结构化 非结构化数据 3](#_Toc14493)

[4. 大数据特点: 价值密度低 3](#_Toc23069)

[5. 大数据的应用: 3](#_Toc25634)

[7.1. 行为分析 3](#_Toc10097)

[8. 大数据优点 实时性 4](#_Toc29344)

[9. 关联技术 4](#_Toc26163)

[9.1. 刷票器/自动化web操作/浏览器引擎 4](#_Toc3782)

[9.2. 注册机/登录器/ 4](#_Toc17175)

[9.3. 发帖机/ 4](#_Toc13371)

[9.4. 语音识别技术 4](#_Toc11769)

[9.5. 手写识别技术 4](#_Toc13038)

[9.6. 验证码../手机验证码. 4](#_Toc112)

[9.7. 水军好评差评 4](#_Toc10745)

[9.8. 云计算、移动互联网 4](#_Toc16798)

[9.9. 物联网. 5](#_Toc7840)

[9.10. 非结构化nosql数据库技术 5](#_Toc14782)

[9.11. 分布式文件系统 5](#_Toc14375)

[9.12. DDBS 分布式存储数据库 5](#_Toc23310)

[9.13. ETL工具 5](#_Toc12194)

[9.14. 采集器/分布式网络爬虫 5](#_Toc29436)

[9.15. 分布式编程语言,ERLANG并行处理计算 5](#_Toc15075)

[9.16. 机器智能化学习 5](#_Toc28092)

[10. 大数据的生产/来源 fromB到fromC 6](#_Toc13362)

[11. 大数据处理的流程(采集>>预处理>>导入>>统计分析>>数据挖掘) 6](#_Toc1984)

[12. 数据采集以及承载 6](#_Toc3190)

[13. 数据清洗、转换、集成ETL工具 6](#_Toc25503)

[14. 大数据的存储 数据库与云存储 文件存储 7](#_Toc17399)

[15. 统计分析 统计分析与模型预测 与结果呈现 7](#_Toc28774)

[16. 大数据的5个基础分析方法可视化分析，数据挖掘，预测 7](#_Toc17892)

[17. 数据挖掘 8](#_Toc11130)

[18. 大数据的传输 8](#_Toc9528)

[19. 应用领域 8](#_Toc25195)

[20. 数据分析师的入门进阶(20天时间) 9](#_Toc28097)

[21. 常用大数据技术 9](#_Toc31841)

[21.1. Erlang分布式大并发编程语言 9](#_Toc4684)

[21.2. 超强的并发性 10](#_Toc20503)

[21.3. 天生的分布式 10](#_Toc2373)

[21.4. 灵活多样的错误处理 10](#_Toc29305)

[21.5. 代码热替换 10](#_Toc29521)

[21.6. ● 软实时性- 11](#_Toc2911)

[22. NoSQL数据库 11](#_Toc17315)

[23. Ddms--Storm —— Twitter开发的ddms 11](#_Toc26879)

[23.1. 4.1. HBase 3 12](#_Toc24609)

[23.2. 4.2. Hypertable 3 12](#_Toc4105)

[23.3. 4.3. Hadoop -----Hive 3 12](#_Toc8087)

[23.4. 4.4. 基于MySQL的分布式数据库实践 3 12](#_Toc13708)

[23.5. 4.5. 使用MySQL federated 引擎构建 MySQL 分布式数据库访问层 3 12](#_Toc32386)

[23.6. 4.6. 用Amoeba构架MySQL分布式数据库环境 4 12](#_Toc14609)

[23.7. 4.7. Facebook开源的Cassandra 4 12](#_Toc26430)

[23.8. 4.8. 淘宝的Amoedb 4 12](#_Toc10391)

[24. 案例: 12](#_Toc4253)

[24.1. 谷歌流感趋势"的工具 12](#_Toc25932)

[24.2. 这个工具工作的原理 13](#_Toc20011)

[24.3. 大数据应用案例之：医疗行业 13](#_Toc30523)

[24.4. 大数据应用案例之：能源行业 13](#_Toc18354)

[24.5. 大数据应用案例之：通信行业 14](#_Toc20299)

[24.6. 大数据应用案例之：零售业 14](#_Toc31159)

[24.7. 麻省理工学院利用手机定位数据和交通数据建立租房规划。[5] [6] 15](#_Toc29242)

[25. 参考: 15](#_Toc12041)

作者Attilax ，  EMAIL:1466519819@qq.com   
来源：attilax的专栏  
地址：http://blog.csdn.net/attilax

# 大数据是什么 4v 大规模 非结构化

大数据(big data)，或称巨量资料，指的是所涉及的资料量规模巨大到无法透过目前主流软件工具，在合理时间内达到撷取、管理、处理、并整理成为帮助企业经营决策更积极目的的资讯。 大数据的4V特点：Volume、Velocity、Variety、Veracity

"大数据"首先是指数据体量(volumes)?大，指代大型数据集，一般在10TB?规模左右，但在实际应用中，很多企业用户把多个数据集放在一起，已经形成了PB级的数据量

数据类别(variety)大，数据来自多种数据源，数据种类和格式日渐丰富，已冲破了以前所限定的结构化数据范畴，囊括了半结构化和非结构化数据。

数据处理速度（Velocity）快，在数据量非常庞大的情况下，也能够做到数据的实时处理。

数据真实性（Veracity）高，随着社交数据、企业内容、交易与应用数据等新数据源的兴趣，传统数据源的局限被打破，企业愈发需要有效的信息之力以确保其真实性及安全性

通过海量搜索词的跟踪获得的趋势报告是很有说服力的，仅波士顿地区，就有700例流感得到确认，该地区目前已宣布进入公共健康紧急状态…点击查看>>

# 大数据技术综述 巨量数据中 快速获得有价值信息的技术

大数据技术是指从各种各样类型的巨量数据中，快速获得有价值信息的技术

# 大数据的格式种类:普通数据 以及图片视频半结构化 非结构化数据

除了普通的文本数据.还有图片,视频,pdf，地理位置信息LBS等等非结构化数据..

视频、图片、半结构化的，如XML文档。

# 大数据特点: 价值密度低

价值密度低。以视频为例，连续不间断监控过程中，可能有用的数据仅仅有一两秒。

# 大数据的应用:

预测,,寻人,解决欺诈侦测、客户获取与维持、网络销售、市场细分、风险分析、亲和性分析、客户满意度、破产预测和投资组合分析

## 行为分析

了解行为的技术会...安全应用.

# 大数据优点 实时性

报告强在“时效性”上，但“大数据”应用突出强调数据处理的实时性

在线个性化推荐

# 关联技术

## 刷票器/自动化web操作/浏览器引擎

## 注册机/登录器/

## 发帖机/

## 语音识别技术

## 手写识别技术

## 验证码../手机验证码.

## 水军好评差评

## 云计算、移动互联网

技术的发展，使得大量数据的生产和连通变成现实；

## 物联网.

## 非结构化nosql数据库技术

非结构化数据库技术的发展，使得数据收集的要求大大降低；

## 分布式文件系统

## DDBS 分布式存储数据库

一个分布式数据库在逻辑上是一个统一的整体，在物理上则是分别存储在不同的物理节点上。一个应用程序通过网络的连接可以访问分布在不同地理位置的数据库。它的分布性表现在数据库中的数据不是存储在同一场地。更确切地讲，不存储在同一计算机的存储设备上。这就是与集中式数据库的区别。从用户的角度看，一个分布式数据库系统在逻辑上和集中式数据库系统一样，用户可以在任何一个场地执行全局应用。就好像那些数据是存储在同一台计算机上，有单个数据库管理系统（DBMS)管理一样，用户并没有什么感觉不一样。

## ETL工具

## 采集器/分布式网络爬虫

NUtch等...

## 分布式编程语言,ERLANG并行处理计算

，使得数据可以得到高速处理，更快获得结果、应用；

## 机器智能化学习

等等又进一步促进大数据的应用发展。所以，我们可以做到存储处理所有数据，而不是存储抽样数据，并且可以将粒度从整体面向个体。这些也带来一系列变革：

# 大数据的生产/来源 fromB到fromC

据创造的主体由企业逐渐转向个体，而个体所产生的绝大部分数据为图片、文档、视频等非结构..语音..

搜索引擎：跟踪搜索词得到相关数据

微博:搜索

Bbs:

微信:

社交SNS数据

电子商务交易纪录

# 大数据处理的流程(采集>>预处理>>导入>>统计分析>>数据挖掘)

　具体的大数据处理方法确实有很多，但是根据笔者长时间的实践，总结了一个普遍适用的大数据处理流程，并且这个流程应该能够对大家理顺大数据的处理有所帮助。整个处理流程可以概括为四步，分别是采集、导入和预处理、统计和分析，最后是数据挖掘。

# 数据采集以及承载

物联网、云计算、移动互联网、车联网、手机、平板电脑、PC以及遍布地球各个角落的各种各样的传感器，无一不是数据来源或者承载的方式。

大数据的采集过程中，其主要特点和挑战是并发数高，因为同时有可能会有成千上万的用户来进行访问和操作，

# 数据清洗、转换、集成ETL工具

虽然采集端本身会有很多数据库，但是如果要对这些海量数据进行有效的分析，还是应该将这些来自前端的数据导入到一个集中的大型分布式数据库，或者分布式存储集群，并且可以在导入基础上做一些简单的清洗和预处理工作。也有一些用户会在导入时使用来自Twitter的Storm来对数据进行流式计算，来满足部分业务的实时计算需求。  
  
　　导入与预处理过程的特点和挑战主要是导入的数据量大，每秒钟的导入量经常会达到百兆，甚至千兆级别。

ETL工具负责将分布的、异构数据源中的数据如关系数据、平面数据文件等抽取到临时中间层后进行清洗、转换、集成，最后加载到数据仓库或数据集市中，成为联机分析处理、数据挖掘的基础。

# 大数据的存储 数据库与云存储 文件存储

**数据存取：**关系数据库、NOSQL、SQL等。

**基础架构：**云存储、分布式文件存储等。

# 统计分析 **统计分析与模型预测 与结果呈现**

**统计分析：**假设检验、显著性检验、差异分析、相关分析、T检验、方差分析、卡方分析、偏相关分析、距离分析、回归分析、简单回归分析、多元回归分析、逐步回归、回归预测与残差分析、岭回归、logistic回归分析、曲线估计、因子分析、聚类分析、主成分分析、因子分析、快速聚类法与聚类法、判别分析、对应分析、多元对应分析（最优尺度分析）、bootstrap技术等等。

**模型预测：**预测模型、机器学习、建模仿真。  
**结果呈现：**云计算、标签云、关系图等。

统计与分析主要利用分布式数据库，或者分布式计算集群来对存储于其内的海量数据进行普通的分析和分类汇总等，以满足大多数常见的分析需求，在这方面，一些实时性需求会用到EMC的GreenPlum、Oracle的Exadata，以及基于MySQL的列式存储Infobright等，而一些批处理，或者基于半结构化数据的需求可以使用Hadoop。  
  
　　统计与分析这部分的主要特点和挑战是分析涉及的数据量大，其对系统资源，特别是I/O会有极大的占用

# 大数据的5个基础分析方法可视化分析，数据挖掘，预测

可视化分析(报表??)，因为可视化分析能够直观的呈现大数据特点，同时能够非常容易被读者所接受，就如同看图说话一样简单明了。

数据挖掘算法，各种数据挖掘的算法基于不同的数据类型和格式才能更加科学的呈现出数据本身具备的特点，

预测性分析能力,大数据分析最终要的应用领域之一就是预测性分析，从大数据中挖掘出特点，通过科学的建立模型，之后便可以通过模型带入新的数据，从而预测未来的数据。

语义引擎:

数据质量和数据管理

# 数据挖掘

与前面统计和分析过程不同的是，数据挖掘一般没有什么预先设定好的主题，主要是在现有数据上面进行基于各种算法的计算，从而起到预测（Predict）的效果，从而实现一些高级别数据分析的需求。比较典型算法有用于聚类的Kmeans、用于统计学习的SVM和用于分类的NaiveBayes，主要使用的工具有Hadoop的Mahout等。该过程的特点和挑战主要是用于挖掘的算法很复杂，并且计算涉及的数据量和计算量都很大，常用数据挖掘算法都以单线程为主。  
  
　　整个大数据处理的普遍流程至少应该满足这四个方面的步骤，才能算得上是一个比较完整的大数据处理

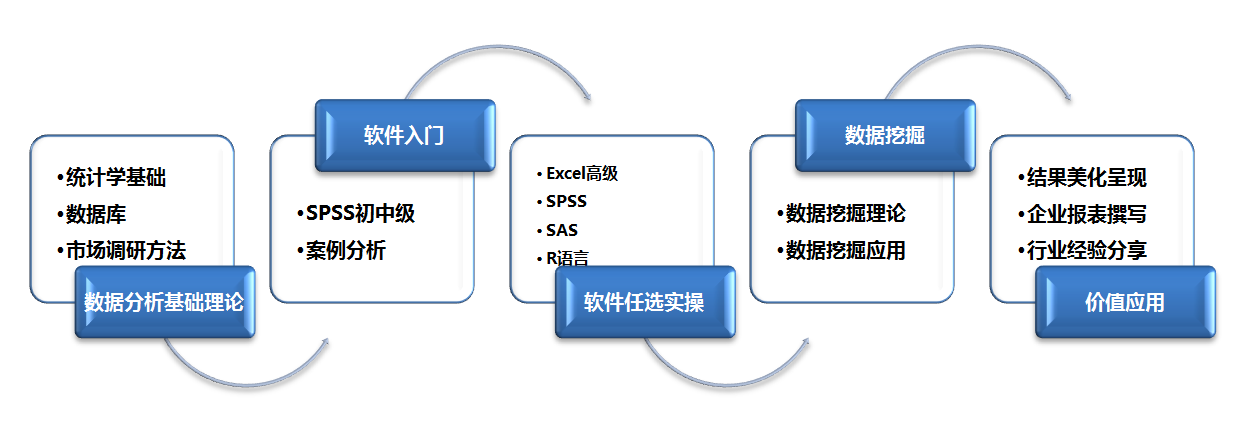
**数据挖掘：**分类 （Classification）、估计（Estimation）、预测（Prediction）、相关性分组或关联规则（Affinity grouping or association rules）、聚类（Clustering）、描述和可视化、Description and Visualization）、复杂数据类型挖掘(Text, Web ,图形图像，视频，音频等)

# 大数据的传输

# 应用领域

所有的领域

# 数据分析师的入门进阶(20天时间)



**，从理论基础、软件应用、数据挖掘到数据应用系统，18天玩转数据分析~**

# 常用大数据技术

## Erlang分布式大并发编程语言

Erlang是一个结构化，动态类型编程语言，内建并行计算支持。最初是由爱立信专门为通信应用设计的，比如控制交换机或者变换协议等，因此非常适 合于构建[分布式](http://baike.baidu.com/view/402382.htm)，实时软并行计算系统。

使用Erlang编写出的应用运行时通常由成千上万个轻量级进程组成，并通过消息传递相互通讯。进程间上下文切换对于Erlang来说仅仅 只是一两个环节，比起C程序的线程切换要高效得多得多了。

使用Erlang来编写[分布式应用](http://baike.baidu.com/view/779017.htm)要简单的多，因为它的分布式机制是透明的：对于程序来说并不知道自己是在分布式运行。

● [并发性](http://baike.baidu.com/view/630151.htm) - Erlang支持超大量级的并发进程，并且不需要[操作系统](http://baike.baidu.com/view/880.htm)具有并发机制..号称一台服务器百万并发..

## 超强的并发性

由于采用其自身Process，而没有采用操作系统的进程和线程，我们可以创建大规模的并发处理，同时还简化了我们的编程复杂度。我们可以通过几十行代码实现一个并发的TCP服务器，这在其他语言中都想都不敢想

## 天生的分布式

Erlang天生适合分布式应用开发，其很多的BIF（内建函数，相API）都具有分布式版本，我们可以通过BIF在远程机器上创建Process，可以向远程机器上的某个Process发送消息。在分布式应用的开发中，我们可以像C、C＋＋，JAVA等语言一样，通过Socket进行通讯，也可以使用Erlang内嵌的基于Cookie的分布式架构，进行开发。当然也可以两者混合。分布式开发更加方便，快速。Erlang的Process的操作，Error的处理等都对支持分布式操作。

● [健壮性](http://baike.baidu.com/view/963152.htm) - Erlang具有多种基本的错误检测能力，它们能够用于构建容错系统。

## 灵活多样的错误处理

Erlang最初为电信产品的开发，这样的目的，决定了其对错误处理的严格要求。Erlang中提供一般语言所提供的exception，catch，try…catch等语法，同时Erlang支持Link和Monitor两种机制，我们可以将Process连接起来，让他们组成一个整体，某个Process出错，或推出时，其他Process都具有得知其推出的能力。而Monitor顾名思义，可以用来监控某个Process，判断其是否退出或出错。所有的这些Erlang都提供内在支持，我们快速的开发坚固的产品，不在是奢望。

## 代码热替换

你的产品想不间断的更新么？Erlang可以满足你这个需求，Erlang会在运行时自动将旧的模块进行替换。一切都静悄悄。

● 热代码升级-Erlang允许程序代码在运行系统中被修改。旧代码能被逐步淘汰而后被新代码替换。在此过渡期间，新旧代码是共存的。

## ● 软实时性-

Erlang支持可编程的“软”[实时系统](http://baike.baidu.com/view/344431.htm)，使用了递增式垃圾收集技术。

# NoSQL数据库

# Ddms--Storm —— Twitter开发的ddms

Storm —— Twitter开发，通常被比作“实时的Hadoop”。**然而Storm远比Hadoop来的简单，因为用它处理大数据不会带来新老技术的交替。**

　对比Hadoop的批处理，Storm是个实时的、分布式以及具备高容错的计算系统。同Hadoop一样Storm也可以处理大批量的数据，然而Storm在保证高可靠性的前提下还可以让处理进行的更加实时;也就是说，所有的信息都会被处理。S

## [4.1. HBase 3](http://blog.csdn.net/attilax/article/details/12851431#_Toc14696)

## [4.2. Hypertable 3](http://blog.csdn.net/attilax/article/details/12851431#_Toc12835)

## [4.3. Hadoop -----Hive 3](http://blog.csdn.net/attilax/article/details/12851431#_Toc26495)

## [4.4. 基于MySQL的分布式数据库实践 3](http://blog.csdn.net/attilax/article/details/12851431#_Toc11059)

## [4.5. 使用MySQL federated 引擎构建 MySQL 分布式数据库访问层 3](http://blog.csdn.net/attilax/article/details/12851431#_Toc11636)

## [4.6. 用Amoeba构架MySQL分布式数据库环境 4](http://blog.csdn.net/attilax/article/details/12851431#_Toc5258)

## [4.7. Facebook开源的Cassandra 4](http://blog.csdn.net/attilax/article/details/12851431#_Toc21482)

## [4.8. 淘宝的Amoedb 4](http://blog.csdn.net/attilax/article/details/12851431#_Toc15698)

# 案例:

## 谷歌流感趋势"的工具

歌有一个名为"谷歌流感趋势"的工具，它通过跟踪搜索词相关数据来判断全美地区的流感情况（比如患者会搜索流感两个字）。近日，这个工具发出警告，全美的流感已经进入"紧张"级别。它对于健康服务产业和流行病专家来说是非常有用的，因为它的时效性极强，能够很好地帮助到疾病暴发的跟踪和处理。事实也证明，通过海量搜索词的跟踪获得的趋势报告是很有说服力的，仅波士顿地区，就有700例流感得到确认，该地区目前已宣布进入公共健康紧急状态[…点击查看>>](http://bbs.pinggu.org/thread-2239985-1-1.html)

## 这个工具工作的原理

大致是这样的：设计人员置入了一些关键词（比如温度计、流感症状、肌肉疼痛、胸闷等），只要用户输入这些关键词，系统就会展开跟踪分析，创建地区流感图表和流感地图。谷歌多次把测试结果（蓝线）与美国疾病控制和预防中心的报告（黄线）做比对，从下图可知，两者结论存在很大相关性：

## 大数据应用案例之：医疗行业

**[1]** Seton Healthcare是采用IBM最新沃森技术医疗保健内容分析预测的首个客户。该技术允许企业找到大量病人相关的临床医疗信息，通过大数据处理，更好地分析病人的信息。  
　　**[2]** 在加拿大多伦多的一家医院，针对早产婴儿，每秒钟有超过3000次的数据读取。通过这些数据分析，医院能够提前知道哪些早产儿出现问题并且有针对性地采取措施，避免早产婴儿夭折。  
　　**[3]** 它让更多的创业者更方便地开发产品，比如通过社交网络来收集数据的健康类App。也许未来数年后，它们搜集的数据能让医生给你的诊断变得更为精确，比方说不是通用的成人每日三次一次一片，而是检测到你的血液中药剂已经代谢完成会自动提醒你再次服药。

## 大数据应用案例之：能源行业

　　[1]　 智能电网现在欧洲已经做到了终端，也就是所谓的智能电表。在德国，为了鼓励利用太阳能，会在家庭安装太阳能，除了卖电给你，当你的太阳能有多余电的时候还可以买回来。通过电网收集每隔五分钟或十分钟收集一次数据，收集来的这些数据可以用来预测客户的用电习惯等，从而推断出在未来2~3个月时间里，整个电网大概需要多少电。有了这个预测后，就可以向发电或者供电企业购买一定数量的电。因为电有点像期货一样，如果提前买就会比较便宜，买现货就比较贵。通过这个预测后，可以降低采购成本。  
  
　　[2]　 维斯塔斯风力系统，依靠的是BigInsights软件和IBM超级计算机，然后对气象数据进行分析，找出安装风力涡轮机和整个风电场最佳的地点。利用大数据，以往需要数周的分析工作，现在仅需要不足1小时便可完成。

## 大数据应用案例之：通信行业

　　[1] XO Communications通过使用IBM SPSS预测分析软件，减少了将近一半的客户流失率。XO现在可以预测客户的行为，发现行为趋势，并找出存在缺陷的环节，从而帮助公司及时采取措施，保留客户。此外，IBM新的Netezza网络分析加速器，将通过提供单个端到端网络、服务、客户分析视图的可扩展平台，帮助通信企业制定更科学、合理决策。  
　　[2] 电信业者透过数以千万计的客户资料，能分析出多种使用者行为和趋势，卖给需要的企业，这是全新的资料经济。  
　　[3] 中国移动通过大数据分析，对企业运营的全业务进行针对性的监控、预警、跟踪。系统在第一时间自动捕捉市场变化，再以最快捷的方式推送给指定负责人，使他在最短时间内获知市场行情。  
　　[4] NTT docomo把手机位置信息和互联网上的信息结合起来，为顾客提供附近的餐饮店信息，接近末班车时间时，提供末班车信息服务。

## 大数据应用案例之：零售业

　　[1] "我们的某个客户，是一家领先的专业时装零售商，通过当地的百货商店、网络及其邮购目录业务为客户提供服务。公司希望向客户提供差异化服务，如何定位公司的差异化，他们通过从 Twitter 和 Facebook 上收集社交信息，更深入的理解化妆品的营销模式，随后他们认识到必须保留两类有价值的客户：高消费者和高影响者。希望通过接受免费化妆服务，让用户进行口碑宣传，这是交易数据与交互数据的完美结合，为业务挑战提供了解决方案。"Informatica的技术帮助这家零售商用社交平台上的数据充实了客户主数据，使他的业务服务更具有目标性。  
　　[2] 零售企业也监控客户的店内走动情况以及与商品的互动。它们将这些数据与交易记录相结合来展开分析，从而在销售哪些商品、如何摆放货品以及何时调整售价上给出意见，此类方法已经帮助某领先零售企业减少了17%的存货，同时在保持市场份额的前提下，增加了高利润率自有品牌商品的比例。

## 麻省理工学院利用手机定位数据和交通数据建立租房规划。[5] [6]

# 参考:

窥视互联网金融：谈谈大数据

Hadoop并非完美：8个代替 HDFS 的绝佳方案