

大数据调研

目录

1	大数据基本概念.....	5
1.1	基本定义与功能.....	5
1.1.1	发展历史.....	5
1.1.2	定义及意义.....	8
1.1.3	主要特点.....	8
1.2	大数据与云计算的关系.....	9
1.2.1	简述.....	9
1.2.2	从存储、处理和分析角度看大数据和云计算的区别.....	9
1.2.2.1	数据存储层.....	10
1.2.2.2	数据处理层.....	11
1.2.2.3	数据分析层.....	11
1.3	技术盘点.....	12
1.3.1	HadoopMapReduce.....	12
1.3.2	NoSQL 数据库.....	12
1.3.3	内存分析.....	12
1.3.4	集成设备.....	13
1.4	大数据研究.....	13
1.5	处理工具.....	13
1.5.1	开源大数据生态圈.....	14
1.5.2	商用大数据生态圈：.....	14
1.6	处理流程.....	14
1.6.1	采集.....	14
1.6.2	导入/预处理.....	15
1.6.3	统计/分析.....	15
1.6.4	挖掘.....	16
2	大数据分析基本概念.....	17
2.1	基本内涵.....	17

2.2	常用工具介绍.....	17
2.3	发展状况.....	18
2.4	应用案例.....	18
3	大数据接地气.....	20
3.1	迁徙地图背后的大数据可视化.....	21
3.2	称赞与吐槽.....	22
3.3	大数据需要更接地气.....	23
4	大数据必备十大工具.....	25
4.1	Apache Hive:	25
4.2	Jaspersoft BI 套件.....	25
4.3	1010data:.....	26
4.4	Actian:.....	26
4.5	Pentaho Business Analytics:	26
4.6	Karmasphere Studio and Analyst:.....	26
4.7	Cloudera:.....	27
4.8	HP Vertica Analytics Platform Version 7:.....	27
4.9	Talend Open Studio:.....	27
4.10	Apache Spark.....	27

说明

本文档含数据挖掘、数据分析、安全分析相关概念、方法与理论。

1 大数据基本概念

1.1 基本定义与功能

大数据技术(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通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。在维克托·迈尔-舍恩伯格及肯尼斯·库克耶编写的《大数据时代》[2] 中大数据指不用随机分析法(抽样调查)这样的捷径, 而采用所有数据进行分析处理。大数据的 4V 特点: Volume(大量)、Velocity(高速)、Variety(多样)、value(价值)。

1.1.1 发展历史

“大数据”这个术语最早期的引用可追溯到 apache org 的开源项目 Nutch。当时, 大数据用来描述为更新网络搜索索引需要同时进行批量处理或分析的大量数据集。随着谷歌 MapReduce 和 Google File System (GFS) 的发布, 大数据不再仅用来描述大量的数据, 还涵盖了处理数据的速度。

早在 1980 年, 著名未来学家阿尔文·托夫勒便在《第三次浪潮》一书中, 将大数据热情地赞颂为“第三次浪潮的华彩乐章”。不过, 大约从 2009 年开始, “163 大数据”才成为互联网信息技术行业的流行词汇。美国互联网数据中心指出, 互联网上的数据每年将增长 50%, 每两年便将翻一番, 而目前世界上 90% 以上的数据是最近几年才产生的。此外, 数据又并非单纯指人们在互联网上发布的信息, 全世界的工业设备、汽车、电表上有着无数的数码传感器, 随时测量和传递着有关位置、运动、震动、温度、湿度乃至空气中化学物质的变化, 也产生了海量的数据信息。

数据革命 - 日益增长的大型传感器、数码设备、企业数据库, 和社交媒体网站 - 改变了一切, 仅仅过去两年就新增了 90% 的数据。从营销人员到政策制定者都已开始接纳诸如大规模数据集和大数据之类松散的定义了。

- 1887 - 1890

美国统计学家赫尔曼·霍尔瑞斯为了统计 1890 年的人口普查数据发明了一台电动机来读取卡片上的洞数, 该设备让美国用一年时间就完成了原本耗时 8 年

的人口普查活动，由此在全球范围内引发了数据处理的新纪元。

- 1935-1937

美国总统富兰克林·罗斯福利用社会保障法开展了美国政府最雄心勃勃的一项数据收集项目，IBM 最终赢得竞标，即需要整理美国的 2600 万个员工和 300 万个雇主的记录。共和党总统候选人阿尔夫兰登 scoffs 嘲笑地说，“要整理如此繁多的职工档案，还必须而调用大规模的现场调查人员去核实那些信息不完整的人员记录。”

- 1943 年

一家英国工厂为了破译二战期间的纳粹密码，让工程师开发了系列开创性的能进行大规模数据处理的机器，并使用了第一台可编程的电子计算机进行运算。该计算机被命名为“巨人”，为了找出拦截信息中的潜在模式，它以每秒钟 5000 字符的速度读取纸卡——将原本需要耗费数周时间才能完成的工作量压缩到了几个小时。破译德国部队前方阵地的信息以后，帮助盟军成功登陆了诺曼底。

- 1997 年

美国宇航局研究员迈克尔·考克斯和大卫·埃尔斯沃斯首次使用“大数据”这一术语来描述 20 世纪 90 年代的挑战：超级计算机生成大量的信息——在考克斯和埃尔斯沃斯按案例中，模拟飞机周围的气流——是不能被处理和可视化的。数据集通常之大，超出了主存储器、本地磁盘，甚至远程磁盘的承载能力。”他们称之为“大数据问题。”

- 2002 年

在 9/11 袭击后，美国政府为阻止恐怖主义已经涉足大规模数据挖掘。前国家安全顾问约翰·波因德克斯特领导国防部整合现有政府的数据集，组建一个用于筛选通信、犯罪、教育、金融、医疗和旅行等记录来识别可疑人的大数据库。一年后国会因担忧公民自由权而停止了这一项目。

- 2004 年

9/11 委员会呼吁反恐机构应统一组建“一个基于网络的信息共享系统”，以便能快处理应接不暇的数据。到 2010 年，美国国家安全局的 30000 名员工将拦截和存储 17 亿年电子邮件、电话和其它通讯日报。与此同时，零售商积累关于客户购物和个人习惯的大量数据，沃尔玛自吹已拥有一个容量为 460 字节的缓存器——

一比当时互联网上的数据量还要多一倍。

- 2007 - 2008

随着社交网络的激增，技术博客和专业人士为“大数据”概念注入新的生机。“当前世界范围内已有的一些其他工具将被大量数据和应用算法所取代”。

《连线》的克里斯·安德森认为当时处于一个“理论终结时代”。一些政府机构和美国的顶尖计算机科学家声称，“应该深入参与大数据计算的开发和部署工作，因为它将直接有利于许多任务的实现。”

- 2009 年 1 月

印度政府建立印度唯一的身份识别管理局，对 12 亿人的指纹、照片，和虹膜进行扫描，并为每人分配 12 位的数字 ID 号码，将数据汇集到世界最大的生物识别数据库中。官员们说它将会起到提高政府的服务效率和减少腐败行为的作用，但批评者担心政府会针对个别人进行剖面分析与分享这些人的私密生活细节。

- 2009 年 5 月

美国总统巴拉克·奥巴马政府推出 data.gov 网站作为政府开放数据计划的部分举措。该网站的超过 4.45 万量数据集被用于保证一些网站和智能手机应用程序来跟踪从航班到产品召回再到特定区域内失业率的信息，这一行动激发了从肯尼亚到英国范围内的政府们相继推出类似举措。

- 2009 年 7 月

应对全球金融危机，联合国秘书长潘基文承诺创建警报系统，抓住“实时数据带给贫穷国家经济危机的影响”。联合国全球脉冲项目已研究了如何利用手机和社交网站的数据源来分析预测从螺旋价格到疾病爆发之类的问题。

- 2011 年 2 月

扫描 2 亿年的页面信息，或 4 兆兆字节磁盘存储，只需几秒即可完成。IBM 的沃森计算机系统在智力竞赛节目《危险边缘》中打败了两名人类挑战者。后来纽约时报配音这一刻为一个“大数据计算的胜利。”

- 2012 年 3 月

美国政府报告要求每个联邦机构都要有一个“大数据”的策略，作为回应，奥巴马政府宣布一项耗资 2 亿美元的大数据研究与发展项目。国家卫生研究院将一套人类基因组项目的数据集存放在亚马逊的计算机云内，同时国防部也承诺要

开发出可“从经验中进行学习”的“自主式”防御系统。中央情报局局长戴维·彼得雷乌斯将军在发帖讨论阿拉伯之春机构通过云计算收集和分析全球社交媒体信息之事时，不禁惊叹我们已经被自卸卡车倒进了“数字尘土”中。

1.1.2 定义及意义

对于“大数据”（Big data）研究机构 Gartner 给出了这样的定义。“大数据”是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。

大数据技术的战略意义不在于掌握庞大的数据信息，而在于对这些含有意义的数据进行专业化处理。换言之，如果把大数据比作一种产业，那么这种产业实现盈利的关键，在于提高对数据的“加工能力”，通过“加工”实现数据的“增值”。

1.1.3 主要特点

大数据分析相比于传统的数据仓库应用，具有数据量大、查询分析复杂等特点。《计算机学报》刊登的“架构大数据：挑战、现状与展望”一文列举了大数据分析平台需要具备的几个重要特性，对当前的主流实现平台——并行数据库、MapReduce 及基于两者的混合架构进行了分析归纳，指出了各自的优势及不足，同时也对各个方向的研究现状及作者在大数据分析方面的努力进行了介绍，对未来研究做了展望。

大数据的 4 个“V”，或者说特点有四个层面：第一，数据体量巨大。从 TB 级别，跃升到 PB 级别；第二，数据类型繁多。前文提到的网络日志、视频、图片、地理位置信息等等。第三，处理速度快，1 秒定律，可从各种类型的数据中快速获得高价值的信息，这一点也是和传统的数据挖掘技术有着本质的不同。第四，只要合理利用数据并对其进行正确、准确的分析，将会带来很高的价值回报。业界将其归纳为 4 个“V”——Volume（大量）、Variety（多样）、Velocity（高速）、Value（价值）。

从某种程度上说，大数据是数据分析的前沿技术。简言之，从各种各样类型的数据中，快速获得有价值信息的能力，就是大数据技术。明白这一点至关重要，也正是这一点促使该技术具备走向众多企业的潜力。

1.2 大数据与云计算的关系

1.2.1 简述

从技术上看，大数据与云计算的关系就像一枚硬币的正反面一样密不可分。大数据必然无法用单台的计算机进行处理，必须采用分布式架构。它的特色在于对海量数据进行分布式数据挖掘，但它必须依托云计算的分布式处理、分布式数据库和云存储、虚拟化技术。

随着云时代的来临，大数据（Big data）也吸引了越来越多的关注。《著云台》的分析师团队认为，大数据（Big data）通常用来形容一个公司创造的大量非结构化数据和半结构化数据，这些数据在下载至关系型数据库用于分析时会花费过多时间和金钱。大数据分析常和云计算联系到一起，因为实时的大型数据集分析需要像 MapReduce 一样的框架来向数十、数百或甚至数千的电脑分配工作。

大数据需要特殊的技术，以有效地处理大量的容忍经过时间内的数据。适用于大数据的技术，包括大规模并行处理（MPP）数据库、数据挖掘电网、分布式文件系统、分布式数据库、云计算平台、互联网和可扩展的存储系统。

1.2.2 从存储、处理和分析角度看大数据和云计算的区别

关于大数据和云计算的关系人们通常会有误解。而且也会把它们混起来说，分别做一句话直白解释就是：云计算就是硬件资源的虚拟化；大数据就是海量数据的高效处理。

虽然上面的一句话解释不是非常的贴切，但是可以帮助你简单的理解二者的区别。另外，如果做一个更形象的解释，云计算相当于我们的计算机和操作系统，将大量的硬件资源虚拟化之后再行分配使用，在云计算领域目前的老大应该算是 Amazon，可以说为云计算提供了商业化的标准，另外值得关注的还有 VMware（其实从这一点可以帮助你理解云计算和虚拟化的关系），开源的云平台最有活力的就是 Openstack 了；

大数据相当于海量数据的“数据库”，而且通观大数据领域的发展也能看出，当前的大数据处理一直在向着近似于传统数据库体验的方向发展，Hadoop 的产生使我们能够用普通机器建立稳定的处理 TB 级数据的集群，把传统而昂贵的并

行计算等概念一下就拉到了我们的面前，但是其不适合数据分析人员使用（因为 MapReduce 开发复杂），所以 PigLatin 和 Hive 出现了（分别是 Yahoo! 和 facebook 发起的项目，说到这补充一下，在大数据领域 Google、facebook、twitter 等前沿的互联网公司作出了很积极和强大的贡献），为我们带来了类 SQL 的操作，到这里操作方式像 SQL 了，但是处理效率很慢，绝对和传统的数据库的处理效率有天壤之别，所以人们又在想怎样在大数据处理上不只是操作方式类 SQL，而处理速度也能“类 SQL”，Google 为我们带来了 Dremel/PowerDrill 等技术，Cloudera（Hadoop 商业化最强的公司，Hadoop 之父 cutting 就在这里负责技术领导）的 Impala 也出现了。

整体来看，未来的趋势是，云计算作为计算资源的底层，支撑着上层的大数据处理，而大数据的发展趋势是，实时交互式的查询效率和分析能力，借用 Google 一篇技术论文中的话，“动一下鼠标就可以在秒级操作 PB 级别的数据”难道不让人兴奋吗？

在谈大数据的时候，首先谈到的就是大数据的 4V 特性，即类型复杂，海量，快速和价值。IBM 原来谈大数据的时候谈 3V，没有价值这个 V。而实际我们来看 4V 更加恰当，价值才是大数据问题解决的最终目标，其它 3V 都是为价值目标服务。在有了 4V 的概念后，就很容易简化的来理解大数据的核心，即大数据的总体架构包括三层，数据存储，数据分析和数据分析。类型复杂和海量由数据存储层解决，快速和时效性要求由数据处理层解决，价值由数据分析层解决。

数据先要通过存储层存储下来，然后根据数据需求和目标来建立相应的数据模型和数据分析指标体系对数据进行分析产生价值。而中间的时效性又通过中间数据处理层提供的强大的并行计算和分布式计算能力来完成。三层相互配合，让大数据最终产生价值。

1.2.2.1 数据存储层

数据有很多分法，有结构化，半结构化，非结构化；也有元数据，主数据，业务数据；还可以分为 GIS，视频，文件，语音，业务交易类各种数据。传统的结构化数据库已经无法满足数据多样性的存储要求，因此在 RDBMS 基础上增加了两种类型，一种是 hdfs 可以直接应用于非结构化文件存储，一种是 nosql 类数据库，可以应用于结构化和半结构化数据存储。

从存储层的搭建来说，关系型数据库，NoSQL 数据库和 hdfs 分布式文件系统三种存储方式都需要。业务应用根据实际情况选择不同的存储模式，但是为了业务的存储和读取方便性，我们可以对存储层进一步的封装，形成一个统一的共享存储服务层，简化这种操作。从用户来讲并不关心底层存储细节，只关心数据的存储和读取的方便性，通过共享数据存储层可以实现在存储上的应用和存储基础设置的彻底解耦。

1.2.2.2 数据处理层

数据处理层核心解决问题在于数据存储出现分布式后带来的数据处理上的复杂度，海量存储后带来了数据处理上的时效性要求，这些都是数据处理层要解决的问题。

在传统的云相关技术架构上，可以将 hive, pig 和 hadoop-mapreduce 框架相关的技术内容全部划入到数据处理层的能力。原来我思考的是将 hive 划入到数据分析层能力不合适，因为 hive 重点还是在真正处理下的复杂查询的拆分，查询结果的重新聚合，而 mapreduce 本身又实现真正的分布式处理能力。

mapreduce 只是实现了一个分布式计算的框架和逻辑，而真正的分析需求的拆分，分析结果的汇总和合并还是需要 hive 层的能力整合。最终的目的很简单，即支持分布式架构下的时效性要求。

1.2.2.3 数据分析层

最后回到分析层，分析层重点是真正挖掘大数据的价值所在，而价值的挖掘核心又在于数据分析和挖掘。那么数据分析层核心仍然在于传统的 BI 分析的内容。包括数据的维度分析，数据的切片，数据的上钻和下钻，cube 等。

数据分析我只关注两个内容，一个就是传统数据仓库下的数据建模，在该数据模型下需要支持上面各种分析方法和分析策略；其次是根据业务目标和业务需求建立的 KPI 指标体系，对应指标体系的分析模型和分析方法。解决这两个问题基本解决数据分析的问题。

传统的 BI 分析通过大量的 ETL 数据抽取和集中化，形成一个完整的数据仓库，而基于大数据的 BI 分析，可能并没有一个集中化的数据仓库，或者将数据仓库本身也是分布式的了，BI 分析的基本方法和思路并没有变化，但是落地到执行的数据存储和数据处理方法却发生了大变化。

谈了这么多，核心还是想说明大数据两大核心为云技术和 BI，离开云技术大数据没有根基和落地可能，离开 BI 和价值，大数据又变化为舍本逐末，丢弃关键目标。简单总结就是大数据目标驱动是 BI，大数据实施落地式云技术。

1.3 技术盘点

1.3.1 HadoopMapReduce

思维模式转变的催化剂是大量新技术的诞生，它们能够处理大数据分析所带来的 3 个 V 的挑战。扎根于开源社区，Hadoop 已经是目前大数据平台中应用率最高的技术，特别是针对诸如文本、社交媒体订阅以及视频等非结构化数据。除分布式文件系统之外，伴随 Hadoop 一同出现的还有进行大数据集处理 MapReduce 架构。根据权威报告显示，许多企业都开始使用或者评估 Hadoop 技术来作为其大数据平台的标准。

1.3.2 NoSQL 数据库

我们生活的时代，相对稳定的数据库市场中还在出现一些新的技术，而且在未来几年，它们会发挥作用。事实上，NoSQL 数据库在一个广义上派系基础上，其本身就包含了几种技术。总体而言，他们关注关系型数据库引擎的限制，如索引、流媒体和高访问量的网站服务。在这些领域，相较关系型数据库引擎，NoSQL 的效率明显更高。

1.3.3 内存分析

在 Gartner 公司评选的 2012 年十大战略技术中，内存分析在个人消费电子设备以及其他嵌入式设备中的应用将会得到快速的发展。随着越来越多的价格低廉的内存用到数据中心中，如何利用这一优势对软件进行最大限度的优化成为关键的问题。内存分析以其实时、高性能的特性，成为大数据分析时代下的“新宠儿”。如何让大数据转化为最佳的洞察力，也许内存分析就是答案。大数据背景下，用户以及 IT 提供商应该将其视为长远发展的技术趋势。

1.3.4 集成设备

随着数据仓库设备(Data Warehouse Appliance)的出现,商业智能以及大数据分析的潜能也被激发出来,许多企业将利用数据仓库新技术的优势提升自身竞争力。集成设备将企业的数据仓库硬件软件整合在一起,提升查询性能、扩充存储空间并获得更多的分析功能,并能够提供同传统数据仓库系统一样的优势。在大数据时代,集成设备将成为企业应对数据挑战的一个重要利器。

1.4 大数据研究

大数据就是互联网发展到现今阶段的一种表象或特征而已,没有必要神话它或对它保持敬畏之心,在以云计算为代表的技术创新大幕的衬托下,这些原本很难收集和使用的数据开始容易被利用起来了,通过各行各业的不断创新,大数据会逐步为人类创造更多的价值。

其次,想要系统的认知大数据,必须要全面而细致的分解它,从三个层面来展开:

第一层面是理论,理论是认知的必经途径,也是被广泛认同和传播的基线。在这里从大数据的特征定义理解行业对大数据的整体描绘和定性;从对大数据价值的探讨来深入解析大数据的珍贵所在;洞悉大数据的发展趋势;从大数据隐私这个特别而重要的视角审视人和数据之间的长久博弈。

第二层面是技术,技术是大数据价值体现的手段和前进的基石。在这里分别从云计算、分布式处理技术、存储技术和感知技术的发展来说明大数据从采集、处理、存储到形成结果的整个过程。

第三层面是实践,实践是大数据的最终价值体现。在这里分别从互联网的大数据,政府的大数据,企业的大数据和个人的大数据四个方面来描绘大数据已经展现的美好景象及即将实现的蓝图。

1.5 处理工具

当前用于分析大数据的工具主要有开源与商用两个生态圈。

1.5.1 开源大数据生态圈

1、Hadoop HDFS、HadoopMapReduce, HBase、Hive 渐次诞生, 早期 Hadoop 生态圈逐步形成。

2、. Hypertable 是另类。它存在于 Hadoop 生态圈之外, 但也曾经有一些用户。

3、NoSQL, membase、MongoDb

1.5.2 商用大数据生态圈:

1、一体机数据库/数据仓库: IBM PureData (Netezza), Oracle Exadata, SAP Hana 等等。

2、数据仓库: Teradata AsterData, EMC GreenPlum, HP Vertica 等等。

3、数据集市: QlikView、Tableau、以及国内的 Yonghong Data Mart。

1.6 处理流程

周涛博士说: 大数据处理数据时代理念的三大转变: 要全体不要抽样, 要效率不要绝对精确, 要相关不要因果。

具体的大数据处理方法其实有很多, 但是根据长时间的实践, 笔者总结了一个基本的大数据处理流程, 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, 以及挖掘。

1.6.1 采集

定义: 利用多种轻型数据库来接收发自客户端的数据, 并且用户可以通过这些数据库来进行简单的查询和处理工作

特点和挑战: 并发系数高

使用的产品: MySQL, Oracle, HBase, Redis 和 MongoDB 等, 并且这些产品的特点各不相同

大数据的采集是指利用多个数据库来接收发自客户端 (Web、App 或者传感器形式等) 的数据, 并且用户可以通过这些数据库来进行简单的查询和处理工作。

比如，电商会使用传统的关系型数据库 MySQL 和 Oracle 等来存储每一笔事务数据，除此之外，Redis 和 MongoDB 这样的 NoSQL 数据库也常用于数据的采集。

在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，比如火车票售票网站和淘宝，它们并发的访问量在峰值时达到上百万，所以需要在采集端部署大量数据库才能支撑。并且如何在这些数据库之间进行负载均衡和分片的确是需要深入的思考和设计。

1.6.2 导入/预处理

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。

1.6.3 统计/分析

定义：将海量的来自前端的数据快速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，利用分布式技术来对存储于其内的集中的海量数据进行普通的查询和分类汇总等，以此满足大多数常见的分析需求

特点和挑战：导入数据量大，查询涉及的数据量大，查询请求多

使用的产品：InfoBright, Hadoop (Pig 和 Hive), YunTable, SAP Hana 和 Oracle Exadata, 除 Hadoop 以做离线分析为主之外，其他产品可做实时分析。

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。

1.6.4 挖掘

定义：基于前面的查询数据进行数据挖掘，来满足高级别的数据分析需求

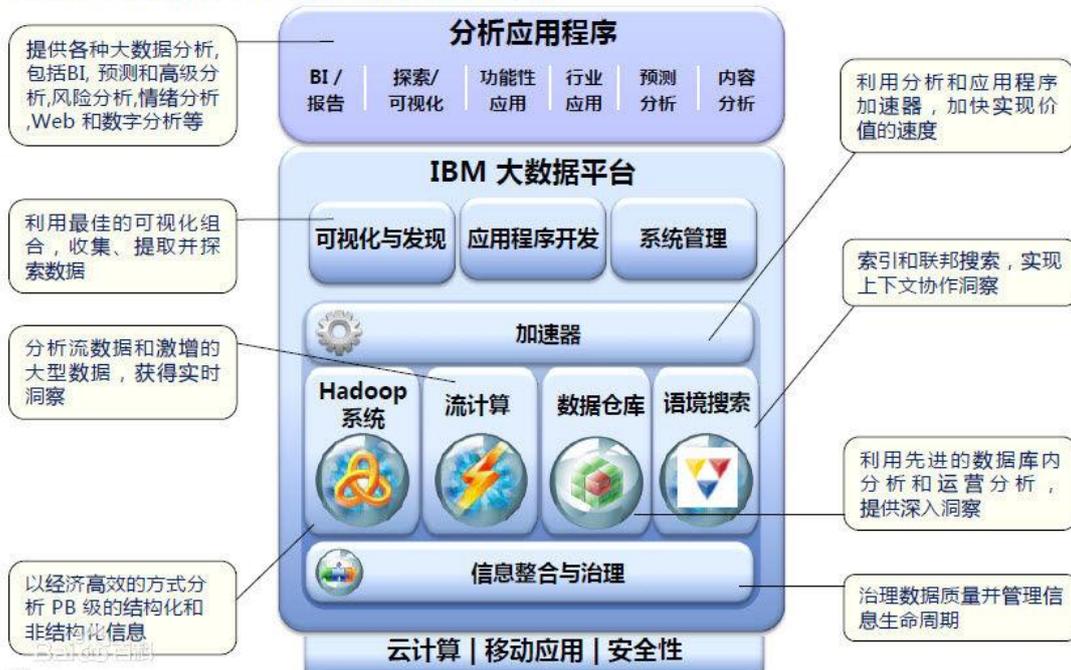
特点和挑战：算法复杂，并且计算涉及的数据量和计算量都大

使用的产品：R, Hadoop Mahout

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 K-Means、用于统计学习的 SVM 和用于分类的 Naive Bayes，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，还有，常用数据挖掘算法都以单线程为主。

IBM 大数据平台和应用程序框架



2 大数据分析基本概念

2.1 基本内涵

大数据分析的五个基本方面

1. Analytic Visualizations (可视化分析)

不管是对数据分析专家还是普通用户，数据可视化是数据分析工具最基本的要求。可视化可以直观的展示数据，让数据自己说话，让观众听到结果。

2. Data Mining Algorithms (数据挖掘算法)

可视化是给人看的，数据挖掘就是给机器看的。集群、分割、孤立点分析还有其他的算法让我们深入数据内部，挖掘价值。这些算法不仅要处理大数据的量，也要处理大数据的速度。

3. Predictive Analytic Capabilities (预测性分析能力)

数据挖掘可以让分析员更好的理解数据，而预测性分析可以让分析员根据可视化分析和数据挖掘的结果做出一些预测性的判断。

4. Semantic Engines (语义引擎)

我们知道由于非结构化数据的多样性带来了数据分析的新的挑战，我们需要一系列的工具有去解析，提取，分析数据。语义引擎需要被设计成能够从“文档”中智能提取信息。

5. Data Quality and Master Data Management (数据质量和数据管理)

数据质量和数据管理是一些管理方面的最佳实践。通过标准化的流程和工具对数据进行处理可以保证一个预先定义好的高质量的分析结果。

假如大数据真的是下一个重要的技术革新的话，我们最好把精力关注在大数据能给我们带来的好处，而不仅仅是挑战。

2.2 常用工具介绍

数据仓库 data warehouse 有：Teradata AsterData, EMC GreenPlum, HP Vertica 等等。

数据集市 data mart 有：QlikView、Tableau、Yonghong Data Mart 等等。

Yonghong Data Mart 是基于自有技术研发的一款数据存储、数据处理的软件。针对客户需要处理需求数据的量级不同，IT 系统架构的不同和存储系统的不同，提供了两种解决方案供客户选择一种本地模式，一种是 MPP 模式。当需要处理的数据量级别处于 TB 级以下，或者采用普通存储结构，或者单机已经足够满足性能需求，建议用户选择本地模式。当面对异构数据库存储系统，需要处理的数据量级别在 TB 级和 PB 级以上，或者 IT 系统和存储系统采用分布式，或者需要 MPP 模式才能满足性能需求，基于分布式架构的并行处理模式更适合客户的需求。

Yonghong Data Mart 底层技术：

1. 分布式计算
2. 分布式通信
3. 内存计算
4. 列存储
5. 库内计算

前端展现：用于展现分析的前端开源工具有 JasperSoft, Pentaho, Spagobi, Openi, Birt 等等。用于展现分析商用分析工具有 Cognos, BO, Microsoft, Oracle, Microstrategy, QlikView、Tableau、国内永洪科技 Yonghong Z-Suite 等等。

2.3 发展状况

开源大数据

(1) Hadoop HDFS、Hadoop MapReduce, HBase、Hive 渐次诞生，早期 Hadoop 生态圈逐步形成。

(2) Hypertable 是另类。它存在于 Hadoop 生态圈之外，但也曾经有一些用户。
一体机数据仓库：IBM PureData(Netezza), Oracle Exadata, SAP Hana 等等。

2.4 应用案例

2014 年 6 月 28 日，奥地利研究人员发表研究公报称，通过对多家网上博彩公司长期以来的赔率、各球队的历史表现和近期球员伤病情况进行大数据分析，他们预测东道主巴西队问鼎世界杯胜算较大。

奥地利因斯布鲁克大学与维也纳经济大学的研究人员推出了一套“博彩共识模型”。根据这套大数据分析模型，巴西队问鼎本届世界杯的几率为 22.5%，阿根廷队为 15.8%，德国队为 13.4%。从数据上看，东道主夺冠的胜算大大超过其他国家队。

3 大数据接地气

每年的这个时候，总会流传着一张图片，那就是非洲的角马大迁徙和春运盛况的对比图，隐含的意思无非就是说“雨季又过了，又到了春运的季节，男男女女们挤在一起，随着列车轻轻的摇动，就如同那雄海龟趴在雌海龟的身上……”



所以当百度推出的春运迁徙地图在央视亮相的时候，马上让我眼前一亮。作为这种大数据可视化产品的脑残粉，一定要跳出来赞一下。

3.1 迁徙地图背后的大数据可视化



从全国迁徙图首页可以看出，数据来源是百度地图 LBS 开放平台，并且辗转找到此项目的负责人求证后，也验证了这一产品的数据来自于众多使用了百度地图的应用所传送来的定位请求，从而对所有请求信息进行辨认设备和定位位置变化来分析处理全样数据。

上面的话比较拗口，简单点儿说，就是只要你的手机里装有使用百度地图 API 的应用，那么你的长距离移动就是这张地图里的一条线。

只拿此刻的数据进行一下解读，1月26日上午十点，在过去八小时内最热的迁入城市前三名是北京重庆和赣州。无论重庆和赣州，都是劳务输出的重点地区，排名前三理所应当。那么北京为什么位居迁入城市第一？

点开北京的路线详情就能看到，迁入北京的大部分是廊坊、天津、葫芦岛等地的人，只是把北京当做一个交通中转站而已。这也就是北京能在迁出城市和迁入城市都能名列第一的原因了。已经看到有人质疑这种产品有什么作用。在我看来，再牛逼的大数据挖掘技术，如果不能以一个接地气的方式表达出来，那么永远就只能停留在拙劣的公关 PR 稿中。举一个最浅显的例子，如果铁道部看完这个图，那么他们至少知道下一步的高铁线路应该怎么铺设。如果你是那个在火车上卖 WIFI 的小哥，你肯定也会选择最热线路吧？卖烧不坏的袜子、越南跌打膏之类的朋友们同理。

前两天，陌陌也推出了他们春运版的数据，盘点了热门回家线路热门群组各种数据。但是看完之后，我陷入了深深的失望，我最盼望的数据是“漂亮妹子最多的线路”“漂亮妹子最多的车厢”“D 杯以上无座只好站着的漂亮妹子最多

的车次” ……

3.2 称赞与吐槽

一直以来，百度是我心目中人格最分裂的公司，A 面是一个善于营销和自我推广的公司，但是最近几年推出的新产品基本都是跟随型产品，别人先蹚出一条血路，然后百度再用自己庞大的用户群和流量去拓宽这条路，包括踩死先行者。B 面是一个拥有着众多牛人和牛逼数据的 GEEK，但是却不拿这些数据来做一些让人拍案叫绝的东西。



如果想成为谷歌一样受到全世界尊重的搜索公司，那么百度必须要做一些让人惊叹的产品。例如 2008 年前，谷歌推出了一个单独的小产品——流感疫情地图，里面将从世界各国卫生组织收集到的流感信息用可视化的方式呈现出来，这样你在出差的时候，就知道是否应该带药品了。（嗯，根据地图显示，我国人民身体素质很好，身体倍棒吃嘛嘛香不得感冒）。



从表面上看，这产品对谷歌商业化产品没有任何拉动作用，但是经过这样的

尝试，在两年后，当 h1N1 病毒肆虐的时候，谷歌已经能将患病高发区整合进自己的地图应用了。

百度迁徙地图算是百度近年来比较少见的，不以拉动任何产品下载使用为目的的数据产品了。但是就产品细节来说，有不少地方有待改进，例如视觉的炫酷感，例如地图的可点击操作等等。



无独有偶，就在五天前，英国《卫报》推出了他们的一个数据产品，叫做《在天上——航空的百年史》。因为 1914 年是世界上首个商业航班试飞成功，当时只有 1 名乘客。1914 年全年也只卖了 1205 张票，而 2013 年卖了 31 亿 2 千万张飞机票。产品首页就实时展示了现在全世界上空正在飞行的飞机数和过去 24 小时所飞过的航线图，很炫很酷，并且密集恐惧症患者慎入。（地址，可能需要翻一下~）

3.3 大数据需要更接地气

大数据，要玩起来，才会更好玩。我们经常会在公关 PR 稿中看到这样的话“在本次发布会上推出的新版本，是基于大数据，由业内资深的大数据挖掘团队和机器学习团队埋头研究数月才推出的……”让人不明觉厉。同样的句式，放之四海而皆准，例如情趣用品，也可以说“我们这次推出的新的按摩棒，是基于大数据，由业内资深大数据挖掘团队和机器学习团队埋头研究数月才推出，完全符合绝大多数中国女性的使用习惯……”

所以，大数据现在需要的是将一个泛概念变成一个个接地气的产品或者项

目。例如美国梅西百货，他们会根据库存和需求变化情况，实时的调整 7300 万种商品的实时定价。例如洛杉矶警局，会根据各个区域之前的犯罪率和居住情况，预测性的调整巡逻频率和力度。

百度迁徙地图，如今只是刚上线，所以更多起到的是公关和宣传的作用。而如果这个产品能坚持 10 年，那么这一定是了解中国产业结构变化和人群生态变化的最简单的途径。

4 大数据必备十大工具

大数据的日益增长,给企业管理大量的数据带来了挑战的同时也带来了一些机遇。



下面是用于信息化管理的大数据工具列表:

4.1 Apache Hive:

Hive 是一个建立在 Hadoop 上的开源数据仓库基础设施,通过 Hive 可以很容易的进行数据的 ETL,对数据进行结构化处理,并对 Hadoop 上大数据文件进行查询和处理等。Hive 提供了一种简单的类似 SQL 的查询语言—HiveQL,这为熟悉 SQL 语言的用户查询数据提供了方便。

4.2 Jaspersoft BI 套件

Jaspersoft 包是一个通过数据库列生成报表的开源软件。行业领导者发现 Jaspersoft 软件是一流的,许多企业已经使用它来将 SQL 表转化为 pdf,, 这使

每个人都可以在会议上对其进行审议。另外，JasperReports 提供了一个连接配置单元来替代 HBase。

4.31010data:

1010data 创立于 2000 年，是一个总部设在纽约的分析型云服务，旨在为华尔街的客户提供服务，甚至包括 NYSE Euronext、游戏和电信的客户。它在设计上支持可伸缩性的大规模并行处理。它也有它自己的查询语言，支持 SQL 函数和广泛的查询类型，包括图和时间序列分析。这个私有云的方法减少了客户在基础设施管理和扩展方面的压力。

4.4Actian:

Actian 之前的名字叫做 Ingres Corp，它拥有超过一万客户而且正在扩增。它通过 Vectorwise 以及对 ParAccel 实现了扩展。这些发展分别导致了 Actian Vector 和 Actian Matrix 的创建。它有 Apache, Cloudera, Hortonworks 以及其他发行版本可供选择。

4.5Pentaho Business Analytics:

从某种意义上说，Pentaho 与 Jaspersoft 相比起来，尽管 Pentaho 开始于报告生成引擎，但它目前通过简化新来源中获取信息的过程来支持大数据处理。Pentaho 的工具可以连接到 NoSQL 数据库，例如 MongoDB 和 Cassandra。Peter Wayner 指出，Pentaho Data(一个更有趣的图形编程界面工具)有很多内置模块，你可以把它们拖放到一个图片上，然后将它们连接起来。

4.6Karmasphere Studio and Analyst:

Karmasphere Studio 是一组构建在 Eclipse 上的插件，它是一个更易于创建和运行 Hadoop 任务的专用 IDE。在配置一个 Hadoop 工作时，Karmasphere 工具将引导您完成每个步骤并显示部分结果。当出现所有数据处于同一个 Hadoop 集群的情况时，Karmasphere Analyst 旨在简化筛选的过程。

4.7 Cloudera:

Cloudera 正在努力为开源 Hadoop, 提供支持, 同时将数据处理框架延伸到一个全面的“企业数据中心”范畴, 这个数据中心可以作为首选目标和管理企业所有数据的中心点。Hadoop 可以作为目标数据仓库, 高效的数据平台, 或现有数据仓库的 ETL 来源。企业规模可以用作集成 Hadoop 与传统数据仓库的基础。Cloudera 致力于成为数据管理的“重心”。

4.8 HP Vertica Analytics Platform Version 7:

HP 提供了用于加载 Hadoop 软件发行版所需的参考硬件配置, 因为它本身并没有自己的 Hadoop 版本。计算机行业领袖将其大数据平台架构命名为 HAVEn(意为 Hadoop, Autonomy, Vertica, Enterprise Security and “n” applications)。惠普在 Vertica 7 版本中增加了一个“FlexZone”, 允许用户在定义数据库方案以及相关分析、报告之前探索大型数据集中的数据。这个版本通过使用 HCatalog 作为元数据存储, 与 Hadoop 集成后为用户提供了一种探索 HDFS 数据表格视图的方法。

4.9 Talend Open Studio:

Talend's 工具用于协助进行数据质量、数据集成和数据管理等方面工作。Talend 是一个统一的平台, 它通过提供一个统一的, 跨企业边界生命周期管理的环境, 使数据管理和应用更简单便捷。这种设计可以帮助企业构建灵活、高性能的企业架构, 在次架构下, 集成并启用百分之百开源服务的分布式应用程序变为可能。

4.10 Apache Spark

Apache Spark 是 Hadoop 开源生态系统的新成员。它提供了一个比 Hive 更快的查询引擎, 因为它依赖于自己的数据处理框架而不是依靠 Hadoop 的 HDFS 服务。同时, 它还用于事件流处理、实时查询和机器学习等方面。