IBM软件

Information Management

驾驭大数据的力量

大数据至关重要的事情

1. 什么是大数据

大数据这个术语有点用词不当。说实话，我们甚至不太喜欢这个术语(尽管事实上，它在本书的封面占据了非常显眼的位置)，因为它暗示着，其他数据有点小(可能是)，或这种特殊类型的数据在良机上比较大(它可能会这样，但并非总是如此)。为此，我们认为最好还是专门用一章来解释大数据究竟是什么。

为什么大数据如此重要?

在尝试了解什么是大数据之前，您应该知道为什么大数据对业务如此重要。简单地说，对大数据的追求直接归因于数据分析，对于企业业务而言，数据分析已经从一个辅助手段发展成为了不可或缺的条件。

事实上，我们应该说对分析的运用水平导致了业界同行的分化: 有些成为领导者，而另一些则成为追随者。很难忽视分析在过去十多年中对组织的影响。IBM商业价值研究院与MIT的斯隆管理评论在一份名为The New Intelligent Enterprise(全新的智能企业)的文章中发布了研究结果。此文章得出的结论是，使用分析获得竞争优势的组织业绩大幅超越其业界同行的可能性在2倍以上。想想看: 分析(具体而言，依赖于大数据所作的分析)将帮助您超越竞争对手，因此，如果您的企业对大数据只是好奇，而竞争对手则不仅仅是好奇，您明白了吧。

大数据的目的就是更好地分析更广泛的数据，并因此代表了在业内同行之间创造更多差异化的机会。这常常是被忽视的一个关键点: 从来没有人通过数据存储中提供一分钱的价值。许多供应商都在谈论大数据，但除了存储大量数据的能力以外，我们并没有看到什么，而要想让这些数据变得有意义，让组织 "部署自己的" 应用程序没有什么帮助。真正的价值只能出现在消费品分析平台上，这种平台让您不必从头开始构建应用程序，有效地将获得洞察的时间曲线拉平。大数据真正的重点是分析。

IBM/MIT在The New Intelligent Enterprise(全新的智能企业)中的联合研究还发现，自一年前进行的上一次研究以来，试图使用分析打造竞争优势的企业的数量猛增了近60%。这项研究的观察结论是，近六成的组织现在通过分析实现了差异化。

简单地说，早期的分析采用者在扩大其领导地位。如果您想成为领导者，就必须充分运用分析，如果您想站在分析的前沿，就必须拥抱大数据。

现在进入“什么是大数据”部分

多年以前，IBM推出了智慧的地球(“物联化、互联化以及智能化”)，这预示了短短几年后对IT格局造成巨大冲击的大数据热潮。

我们认为，沃尔玛对使用无线射频识别(RFID)标签进行供应链优化的推动是一个非常好的故事，它是大数据时代来临的一个例证。RFID是以机器速度生成可被收集和分析的数据的一个出色例子。今天，世界已经变得更加物联化和互联化，这要归功于包括RFID标签在内的很多技术。RFID技术的示例包括，在滑道级别或库存单位(SKU)级别跟踪产品；跟踪实时库存；使用徽章来跟踪参加会议的人员；监测在运输过程中的食品温度；跟踪行李(从我们作为旅客的经验，这方面还有很多改进的余地)；监测桥梁混凝土结构的状况；以及监控铁路轨道的热膨胀度，然后根据该数值调整车速，我们还可以举出数千种其他用例。

在2005年，估计有13亿RFID标签在流通；至2011年年底，这一数字已上升到300亿！现在，考虑到RFID的价格点预计到2015年会下降到低于1美分，并且已经有各种其他传感和测量技术可用；事实上，我们此时会讨论，我们可以测量这个世界上任何想测量的东西。

从物联化的角度来看，如今还有什么不包含一定量的编码吗？看看您的汽车：这些天，不给它连接一台电脑都不能诊断出问题。在如今的硬件网络交换机中，软件代码比组件更多。最新的航空客运飞机仪表与超过十亿行的代码物联，这些代码在每1.5小时的操作中生成每发动机约10兆兆字节(TB)的数据。让我们说得更明白一点：从伦敦的希思罗机场到纽约的约翰·F·肯尼迪机场的一次单程飞行将产生大约650TB数据！这可能比如今您的仓库中的数据还要多。大部分这些数据可能从来都没有被看过，除非有灾难发生。如果经济高效地分析所有这些数据，想象一下所获得的效率、潜在的灾难预防成果、洞察和业务优化等其他机会。

其中一个重要的企业差异化因素是捕获“掉到地板上”的数据的能力；此类数据可以产生令人难以置信的洞察和结果，因为它可以丰富在您的组织中正在执行的分析手段。数据废气(Data exhaust)是我们对此类数据喜欢使用的术语：其数量巨大(通常每天若干TB)，但通常不会深挖它来获得业务洞察。在线店面无法捕获多个TB的点击流来执行Web会话化(sessionization)、优化“最后一公里”的购物体验，并且或许也无法了解为什么在线购物篮被放弃。我们可以通过收集并分析堆积如山的数据来判断一个石油钻井平台的状况。也可以分析您最重要的网络的日志文件，在故障出现之前就提供相应的预测和预警的能力，像“大海捞针”一样找出

可能表示下游问题的指标。此处有一个“如果”与大数据的承诺紧密相关：“如果您能收集和分析所有数据……”我们喜欢把对所有数据进行分析的功能称为整体分析。这是大数据的价值主张之一。此处考虑的是，如果分析程序没有受限于数据的采样和收集，它们可以做出什么预测和洞察。

如今，许多公用事业公司的长期计划中都包括逐渐采用智能电表和电网，确保实现可靠的能源供应，结合分布式发电资源，并让客户能够对其能源使用方式有更多的控制。很多公司的第一步都是部署智能电表系统，这意味着一个直接的技术挑战：从每月一次电表读数，变成每15分钟一次智能电表读数，可换算为每一百万个仪表每天产生9600万次读数：数据收集速率增加了3000倍！您可以想像，如果没有适当的管理，这种数据生成速率可能造成沉重的负担。

当然，它也有好处。额外的数据开辟了新的机会，让能源企业可以做他们以前从来不可能做到的事情。利用智能电表采集的数据可以更好地理解客户的细分和行为，以及预测定价如何影响使用情况，但前提是这些企业有能力使用这些数据。例如，按使用时间定价鼓励精于成本计算的能源消费者在非高峰时间运行其洗衣设施、空调和洗碗机。不过，机会还不止这些。利用智能电表和智能电网提供的更多信息，有可能改造并极大地提高发电和调度的效率。

智能电表是智能的，因为它们不仅可以同消费者沟通电力的使用情况和价格，也可以同公用事业提供商沟通电压波动或断电的确切位置。智能电表在产生大量新信息，从根本上改变了公用事业公司与其客户的互动方式。

还有生产消费者的出现，这是一个全新的消费者类别，他同时也是生产者。生产消费者通过太阳能电池板发电，并将电力卖回给电网；这样做在整个供应链中也产生了连锁反应。通过对其数据进行预测分析，公司可以作出广泛的预测，如在销售和传输方面考虑记录过剩的电力、典型的故障点和电网中断位置，以及哪些客户有可能将电力回馈到电网，他们可能在什么时候这样做。

现在考虑社交媒体的更多影响。在物联化和互联化的世界之上的一个社交层也会产生大量数据。这些数据更复杂，因为它们大部分是非结构化的(图像、Twitter内容、Facebook帖子、微博评论等)。如果您吃过Frito-Lay SunChips，可能还记得它改用世界上第一个可生物降解的环保型薯片袋；您可能还记得这个包装有多大声。客户创建了数千个YouTube视频来显示环保袋有多吵。一个“对不起，这个SunChips袋太吵了，我听不见你说什么”Facebook页面有超过50,000个关注者，但博主让大家知道了自己的感受。最后，Frito-Lay推出了比较安静的一款全新SunChips袋，这证明了社交媒体的力量和重要性。

多年来，Facebook每三秒钟就增加一个新用户，如今，这些用户每天生成两位数TB级的数据。事实上，Facebook一天中一般会有超过25亿个关注和30万次照片上传。Facebook帖子的格式其实是结构化数据；它使用JavaScript Object Notation (JSON)格式进行标记:

{

"data":[

{"id":"53423432999_23423423_19898799",

"from":{"name":"Paul Zikopoulos","id": "Z12"},

"message": "Thinking of surprising my wife with a quality time gift that lets her know she's special, any ideas? I thought about taking her to the driving range, perhaps play a round and caddie my game.",

"created_time": "2012-08-02T21:27: 44+0000", "likes: 5,"

"comments": {"data": [{"id": 2847923942_723423423", "from": { "name": "MaryAnne Infanti", "id": "948574763" },

"message": "Paul! Purses and gold! Costco´ s got a great Kate Spade purse on sale this week that says I love you without having to lift a pen.

If you go with your idea, the only thing driving will be you: alone! ", "created_time 2012-00-02T11:27:44+0000", "likes: 64 } }

这个Facebook帖子是结构化的，虽然这一点是毫无疑问的，但其非结构化的部分才更具有潜在价值；它包含了一个糟糕的计划意图，以及强烈建议可能有什么更好计划的评论。对结构化数据进行存储和分析很容易；但是，要分析其非结构化组件中的意图、情绪等，这是非常困难的，但它有可能产生非常高的回报，如果……

Twitter是另一种现象。这个世界所产生的有关体育赛事、销售、图像、政治等的简短意见(140个字符以内)和评论(通常是未经筛选的)已经达到两位数的TB。Twitter也是另一种媒体，提供了格式是结构化的大量数据，但其结构内的非结构化部分才真正保存了大部分未挖掘的价值。看看Noah Kravitz (@noahkravitz)的例子，在离开其公司跳槽到竞争对手的公司之前，当他在某公司任职时，他有超过25,000名关注者。而他辞职以后，他的前雇主起诉他，声称Kravitz先生的Twitter关注者代表了属于雇主的客户名单(想象一下自己成为一宗法庭诉讼的主角)。该案今天仍未有定论，并且肯定将会成为一个先例，但它说明了在Twitter生态系统中体现的

价值，如果不是事实的价值，至少是感知的价值(我们认为是前者)。

在今天收集到的数据中，大部分在时间和空间上都很丰富。例如，我们知道电视节目Myth Busters的其中一位明星住在哪里——不是因为他告诉过我们，而是因为他在启用了基于位置的服务(LBS)的智能设备上将其汽车的照片发送到Tweeter上，从而与超过650,000位最亲密的朋友共享了他家的地理(纬度/经度)坐标！大部分人都不知道什么是LBS，但他们都打开了它，因为他们都在使用某些移动地图应用程序。现在，通过可将地理坐标转换成容易识别位置的社交应用程序，人们就可以让您知道他们什么时候去健身房，或者他们在哪家餐厅就餐。此类数据往往具有内置的位置感知，这代表了实现更精细粒度的个性化或情况风险评估的另一个巨大机会，如果……如今，一些大型信用卡公司拥有基于这种方法的计划供您加入；例如，如果您使用自己的信用卡购买咖啡，他们会分析您的位置(通过LBS)、购买记录，并提供在您当前位置附近的零售商专门为您量身定制的优惠活动。

时间戳无处不在，包括用您的相机或智能手机拍摄的照片上的自动日期元数据、Facebook帖子的发布时间，您打开智能手机或观看您最喜爱的节目的时间等；其实，建立一个生活时间表变得很容易。如果您试想一下，平均来说，英国伦敦的乘客每天从伦敦市中心回家的路上会拍照150多次，然后将这些照片加上在这段时间框架所产生的各种情绪、时间和空间数据，您就已经获得了可随意处置的很多信息——大数据信息。

通过字母V带给您: 我们如何定义大数据

为了简单起见，我们通常使用四个V来定义大数据，即数量(volume)、种类(variety)、速度(velocity)和真实性(veracity)。我们最近增加了真实性这个特征，旨在响应我们的客户在其大数据项目中开始面临的质量和来源问题。有些分析师还会包括其他基于V的描述符，如变异性(variability)和可见性(visibility)，但我们在本次讨论中不涉及这些方面。

毫无疑问: 数据的数量正在增加

数量是明显的大数据特征。在本章的开始，我们就罗列了各种数量统计资料，这些统计做两件事情: 在引用它们的那一刻就过时了，并且变得更大！我们可以同家庭存储成本联系起来，还记得向朋友吹嘘我们花了500美元买回来的新1TB硬盘，它现在大约卖60美元；再过几年，一个消费品版本的硬盘将只有指甲大小。

关于大数据和数据数量的问题是，语言发生了变化。曾经以petabytes (PB)来衡量的总

计内容，现在要用听起来像是来自“星球大战”电影的术语来形容: zettabytes(ZB)。一个zettabytes是一万亿个gigabytes (GB)，或十亿个terabytes (TB)！

由于我们在上一节已经提供了一些很好的数据数量的示例，本节内容会比较少，并以引用世界总体的数字数据增长速度来结尾。在2009年，全球大约有0.8ZB的数据；在2010年，我们超过了1ZB，而在2011年年底，这个数字估计为1.8ZB(我们认为80%是相当高的增长速率)。从现在开始的六到七年后，该数字估计(注意，本书中对未来的任何估计在我们保存草稿的那一刻已经过时了，并且偏低)是35ZB左右，相当于约四万亿台8GB iPod的容量！考虑到这是一个偏低的估计，这个数字仍是惊人的。同样惊人的是与这个数量的数据相关的挑战和机遇。

种类是生活的调味品

大数据的种类特征的真正目的是尝试捕获决策制定流程涉及的所有数据。让非结构化数据有意义(例如在Facebook上的意见和意图想法)或分析图像，并不是计算机与生俱来的功能。然而，这种数据补充了今天我们用于推动决策的数据。那里的大部分数据都是半结构化或非结构化的。(澄清一下，所有数据都有一定的结构，当我们提到非结构化数据，我们指的是没有结构的子组件，如在评论字段输入的自由格式的文本，或自动记录日期的图片中的图像。)

考虑一个客户呼叫中心；想象一下，如果能够检测到一位愤怒的客户提高声音说“这是我在一个星期中的第三次中断！”中的语调变化，大数据解决方案不仅将“第三次”和“中断”识别为消费者脆弱性的负面趋势，同时也将语调变化作为客户流失事件趋向发生的另一个指标。所有这种洞察都可以从非结构化数据中收集到。结合这种非结构化数据与客户的记录数据和交易历史(我们熟悉的结构化数据)，您就拥有了此客户的一个非常个性化的模型: 他的价值、他作为您的客户的属性等。(您可以从尝试非实时地分析已记录的电话开始这个使用模式，随着时间的推移让解决方案逐渐成熟，最终实现实时地分析口头语言)。

IBM业务合作伙伴TerraEchos已开发了一个世界上最成熟的声音分类系统。该系统用于实时的周边安全控制；有一千个传感器埋在地下，以收集和分类所检测到的声音，从而根据分类采取适当的行动(派遣人员、派遣空中监视等)。试想由绿地包围的核反应堆周边出现安全问题。TerraEchos系统可以近乎瞬间地区分出风声和人声，或者区分出一头鹿跑动的声音和人的脚步声。事实上，如果在其保护的森林中有一棵树要倒下，TerraEchos可以肯定，即使周围没有人听到，它还是发出了声音。声音分类是大数据种类特征的一个不错示例。

您的分析能有多快？数据的速度

我们最喜欢但了解最少的一个大数据特征是速度。我们将速度定义为数据到达企业并被处理或充分了解的速率。事实上，我们请客户问问自己，一旦数据到达其企业的门口："您需要多长时间才可以处理它，或者知道它已经到达？"

仔细想想。数据的机会成本时钟在数据到达线缆的那一刻已开始计时。作为组织，我们发现趋势或者获得宝贵洞察所花的时间太长了。不论您在什么行业，能够更迅速地了解和响应数据信号，都会让您处于主导地位。无论您想了解交通系统的状况、患者的健康，还是贷款组合的状况，更快的反应速度都会给您提供一个优势。速度也许是大数据热潮中最容易被忽视的一个领域，并且我们认为，就所提供的功能和成熟度而言，IBM在该领域中是无与伦比的。

在如风暴一般占领市场的大数据热潮中，每个人都紧盯着静态分析，使用优化的引擎(如IBM PureData System for Analytics背后的Netezza技术或Hadoop)来执行以前不可能实现的，至少不会如此大规模实现的分析工作。虽然这极为重要，但我们必须要问："您如何对运动中的数据进行分析？"此功能有可能为企业提供最高水平的差异化，但它似乎在一定程度上受到了忽视。IBM InfoSphere Streams (Streams) 是IBM大数据平台的一部分，它提供了实时的流式传输数据分析引擎。Streams是一个平台，可对时间序列数据包的连续流进行快速、灵活和可扩展的处理。在第三部分中，我们将深入介绍Streams的详情和功能。

您可能会想，复杂事件处理(CEP，Complex Event Processing)系统可以处理速度问题，虽然表面上看来它们可能适用，但在大数据世界中，它们的缺点非常明显。流处理实现跨不同数据类型的高级分析，具有非常高的消息数据速率和极低的延迟(微秒到秒)。例如，金融服务行业(FSS)的客户每秒分析和关联超过五百万条市场消息，以执行期权交易算法，其平均延迟为30微秒。另一位客户分析每秒超过500,000条Internet协议细节的记录(IPDR)，每天超过60亿IPDR，每年超过4PB的数据，以了解其网络状况的趋势和当前状态。考虑一个企业的网络安全问题。在这个领域，威胁是微秒级的，所以您需要一种能够响应和跟上此速度的技术。但是，您也需要能够快速捕获大量数据，并对其进行分析，以便在网络数据包流过网络基础架构时确定网络数据包上新出现的签名和模式。

最后，从治理的角度来看，考虑大数据分析速度引擎的额外好处：如果您有一个强大的分析引擎，可以在数据流过线缆时将非常复杂的分析应用到数据上，您就可以从这些数据中搜集洞察，而不必存储数据，您可能不必让这些数据受到保留政策的约束，而这可能会为您的IT部门带来巨额节省。

今天的CEP解决方案可以针对最多约每秒数万个消息，延迟为几秒到几分钟。此外，分析大多是基于规则的，仅适用于传统的数据类型(与前面的TerraEchos示例相反)。不要误会我们的意思：CEP有其用武之地，但它的设计点有根本性的差异。CEP是一个面向非程序员的解决方案，供使用简单规则的应用程序分离“复杂的”事件。

请注意，正在谈论大数据速度的人并不多，因为可以做到这一点的厂商并不多，更不用说集成静止技术与实时处理，为企业目前的投资提供规模经济效应。请参阅图1-1，仔细考虑您的公司在利用运动、静止的大数据分析平台时将拥有的竞争优势(第3章将详细介绍IBM大数据平台)。

您可以看到大数据如何流进企业；注意，机会成本时钟在左侧的那一点就开始计时。通过的时间越长，您拥有的潜在竞争优势越少，并且所获得的数据回报(ROD)也越低。我们觉得这个ROD指标将在大数据世界中主导未来的IT环境格局：我们以前谈论投资回报(ROI)，它涉及整个解决方案的投资；但是，在大数据世界中，ROD是一个粒度更细的指标，有助于推动未来的大数据投资。传统上，我们使用静止的解决方案(传统的数据仓库、Hadoop、图形存储等)。图1-1中右侧的T方框表示发现和收获静止数据的分析(在本例中，它是基于文本的情绪分析)。

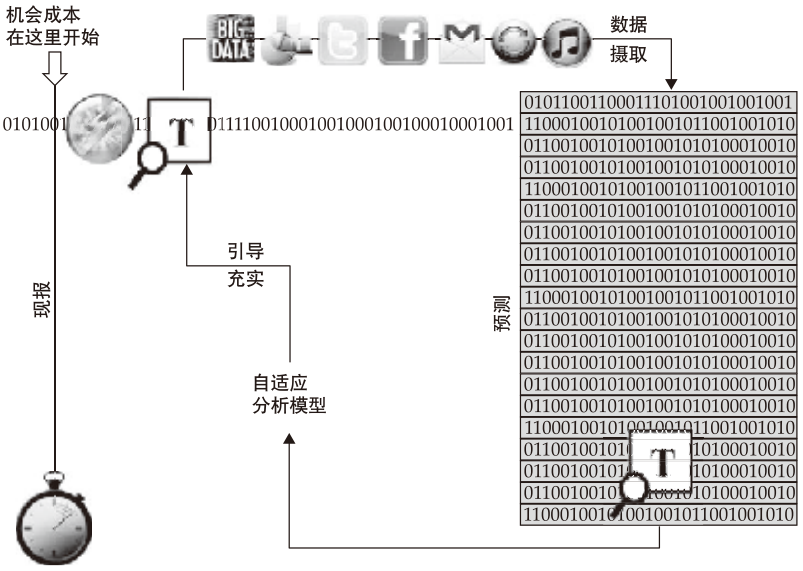


图1-1 实时分析处理(RTAP)生态系统

不幸的是，这是许多供应商的大数据讨论结束的地方。事实上，许多供应商都无法帮助您构建分析；他们只能帮助您执行它。这是您在IBM大数据平台中可以发现的一个关键区别。想象一下，能够无缝地移动在静止数据中获得的分析工件，并在数据运动发生时将该洞察应用到数据(左侧有闪电符号的T方框)。这改变了游戏规则。它让分析模型是自适应的，是有生命的实体，每一天都会变得更聪明，并在数据到达组织的门槛时将所学习到的智能应用到数据。这个模型是周期性的，我们往往将其称为自适应分析，因为这种架构具有实时闭环机制。

对静止数据和运动数据进行无缝分析的能力，让您在与传统仓库(右)紧密匹配的预测模型的基础上有所发展，并通过现报(nowcast)模型激励业务。重点是将您通过静止数据学习到的洞察应用到业务前沿，因此在它发生时可以优化和理解它。出乎意料的是，企业完成此自适应分析周期的时间越长，获得的智能就越多。如果您熟悉在观测的基础上进行调整的控制系统或控制理论，这是一个类似的循环过程。打一个简单的比方，想想当拼图快完成的时候，甚至只有外框架时，要完成它都会变得容易得多。在RTAP系统中，您识别和带给业务前沿的资料更多，所感兴趣的主题画面就会越完整，就会越早在周期中拥有它。

数据在这里，数据在那里，数据，数据无处不在: 数据的真实性

真实性是在描述大数据时使用得越来越多的一个术语；它指的是数据的质量或可信度。协助处理大数据的真实性的工具可将数据转换成可信的洞察并丢弃噪音。

总体来讲，大数据平台让企业有机会分析所有数据(总人口分析)，并更好地了解业务、客户、市场等。这样的机会产生了大数据难题: 虽然删除数据的经济性造成对组织可用的数据量激增，但企业可以理解的数据比例却在下降。更复杂的是，企业试图理解的数据已经饱和，其中包括有用的信号和大量噪声(不能被信任，或者是对于手头的业务问题无用的数据)，如图1-2所示。

我们都亲身体验过这一点；Twitter充满了垃圾机器人和定向推文的例子，它们是不可靠的数据。

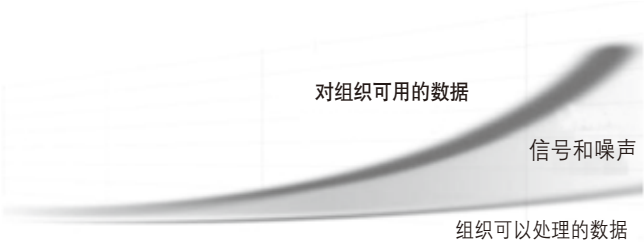


图1-2 随着提供给组织的的数据量在增加，它可以处理的相对数据量在减少。

墨西哥2012年的总统选举最终变成了一个Twitter真实性的示例，各种虚假帐户污染了政治讨论，推出贬义的哈希标签等。垃圾邮件对IT界人士并不陌生，但您需要知道，在大数据世界中，也可能有大垃圾(Big Spam)，您需要有一种方法筛选它，并找出什么数据是可信的和不可信的。当然，还有需要根据上下文、术语等理解的话(我们在第8章讨论这个问题)。

如前所述，有用的信号嵌入在所有这些噪声中：某人自称非常鄙视其目前的智能手机制造商，并开始说需要一台新手机，这个人在表达一种货币化的意图。大数据如此巨大，其质量问题是一个现实，而我们一般用“真实性”来指代这个问题领域。事实上，三分之一的业务领导者不信任自己用于决策的信息，这是一个强烈的信号，一个良好的大数据平台需要解决数据真实性问题。

我的数据仓库在大数据世界中会怎么样?

权威人士坚持认为执行分析的传统方法已过时。有时，这些NoSQL(其实意味着“不仅仅是SQL)权威人士们认为，所有的仓库都会像恐龙一样绝种——考虑到关于NoSQL数据库的许多关注都是如何将SQL接口进行兼容时，这是具有讽刺意味的。没有什么比这离真相更远。我们看到了一些专用的引擎和编程模型非常适合于某些类型的分析。例如，Hadoop的MapReduce编程模型更适用于某些类型的数据，而不是传统的仓库。出于这个原因，您将在第3章中了解到，IBM大数据平台包括一个Hadoop引擎(并支持其他Hadoop引擎，如Cloudera)。更重要的是，IBM认识到编程模型的灵活性，所以IBM PureData System for Analytics(之前被称为Netezza)可以在数据库中执行MapReduce

程序。在大数据时代中真正重要的是，您要选择一个提供非常适合手头任务的专用引擎(您正在执行的分析类型、所分析的数据类型等)的灵活平台。该平台还必须让您能够无缝地跨平台转移编程技能、API和资产，这样，就可以将分析应用到针对手上的数据进行了优化的引擎。例如，IBM大数据平台可以让您采用通过其Annotated Query Language (AQL)构建的文本分析，并无缝地将它们从静止的Hadoop引擎部署到它的Streams大数据实时处理引擎。大部分在Hadoop中编码的MapReduce程序都可以在IBMPureData System for Analytics中运行；在IBM pureSystems for Operational Analytics(之前称为IBM Smart Analytics System)上生成的SQL报告无需修改就可以部署到DB2 for z/OS 上。

考虑数据应存储在哪里时，最好先了解现有数据是如何存储的，以及什么特性是持久性选项的特征。存储在传统数据仓库中的数据在进入仓库前要经过大量的处理。一旦数据到达仓库，就预期它是高质量的，然后通过收集、匹配、转换、元数据、主数据管理、建模和附加到数据上的其他质量服务对数据进行清理，让数据准备好进行分析。显然，这可能是一个成本高昂的过程，而进入仓库的数据被视为同时具有高价值和广泛的用途:它被传输到其他地方，并将出现在以准确性为关键要求的报告和仪表板中。

与此相反，在一些较新的大数据存储库中的数据很少经历(至少在初期)这种严格的预处理，因为这样做的成本将会非常高昂，而且这些存储库中的工作更多地受限于发现，而不是已知的值。更重要的是，每个存储库在不同的应用中有不同的特征需求。有些人可能会优先考虑应用程序的ACID(原子性、一致性、隔离性和持久性)属性，其他人可能以宽松的一致性状态运营，可以容忍BASE属性(基本可用的软状态，并最终达到一致)。

我们喜欢使用黄金开采来表达大数据的机会。在“昔日”(由于某些原因，我们的孩子认为这是我们像他们现在这种年纪的时候)，矿工们可以很容易发现金块或金脉，因为它们是肉眼可见的。让我们把黄金比作“每字节高价值的数据”，您可以看到它的价值，因此，您投入资源来提取它。但是，还有更多的黄金，有可能在附近的山上或数英里之外；肉眼看不见它，试图找到这些隐藏的黄金变得像赌博游戏一样。当然，历史上一直有淘金热的故事，但从来没有人动员数百万人到每个地方都去挖；这样做成本太高了。如今，矿工的工作方式有所不同。黄金开采利用大规模的资本设备，可以处理数百万吨的泥土(每字节低价值的数据)，以发现几乎看不见的金线(黄金的矿石等级通常需要达到30 ppm才能用肉眼观察到)。换句话说，在这些泥土(每字节低价值数据)中有大量的黄金(每字节高价值数据)，使用合适的设备，

就可以经济地处理大量泥土，保留所发现的金片。金片被处理(也许在数据仓库或其他洞察引擎中处理)，熔为金条，最终被存储和记录在一个安全、接受治理、有价值和可信的地方。黄金行业的工作是进行化学清洗，其目的是发现粒度更细的黄金，从之前提取的泥土中找到更多价值(现在想想数据)。我们认为这个比喻非常适合我们的大数据故事，因为我们愿意打赌，如果您有一个由十年的交易数据组成的资料库，与目前使用的技术相比，新的分析方法可以让您在从现在算起三年的数据中提取更多的洞察。

此外，如果您看看作为数据仓库和Hadoop存储库的特征的访问模式，就会发现一个差别，数据仓库的特征往往是让您能够与系统进行交互工作并保证响应时间。事实上，如“思想速度的响应时间”等术语暂时并不是与批处理系统有关的描述，目前是和Hadoop有关。

大数据平台让您可以使用其原生业务对象格式存储所有数据，并通过产品组件上的大规模并行处理从中获得价值。对于交互式查询需求，您要继续选择来源，清洗数据，并将它保存在仓库中。但是您可以通过拉入似乎无关的信息，形成更可靠的视图，从大量低保真数据中得到更多价值。换句话说，数据可以在Hadoop中停留一会儿，当其值经过验证并且可持续时，可以将它迁移到仓库中。

观察和发现之间的差异并不像我们在这里所描述的那样黑白分明。例如，Hadoop引擎的一个常见用例是向数据仓库提供数据库归档服务，将不再“暖”或“热”的数据移到由Hadoop支持的成本较低的存储平台。例如，客户档案系统保存两年以内的高温数据，但在20年业务推移的过程中收集到的所有数据可能仍有价值。保险公司可以受益于了解您的档案，看看您如何从单身发展为已婚，然后又有了孩子，保险公司同时还会考虑目前的趋势或事件(尤其是理财产品组合)。

当然，查询API的可移植性在这种场景中很关键；例如，在访问被迁移到Hadoop的冷数据(IBM大数据平台可以让您做到这一点)时，不必重新编写基于SQL的应用程序代码。另一个示例，也是大数据平台的要求，是SQL和NoSQL世界之间的集成。在您的关系型仓库中可能有一个消费者脆弱性作业在运行，但您选择启动一个基于Hadoop的品牌情绪作业，它可能会影响最后的脆弱性评估。在同一时间，您可能在运行对TB级的点击流日志数据进行分析的Hadoop作业，并且想从记录系统中提取采购信息，以了解还有什么其他因素导致人们成功购买或放弃过去的在线购物车中的商品。(现在试想一下，如果您可以在人们将商品放入购物车时将这种逻辑应用到购物车)。

传统引擎和新的数据处理引擎(Hadoop和其他引擎)将成为一个组织的左膀右臂。关键是确保您的大数据平台供应商可以提供支持同时使用这对臂膀的集成技术。

最后一个比喻，想想棒球运动员。一位典型的棒球运动员非常强壮，用一只手投球，用另一只手接球；大脑协调四肢的活动，以获得最佳的战绩。如果一位棒球手试着使用他非惯用的手来投球或接球，他也许能够做到这一点，但不会很流畅、干脆，看起来也不会非常专业。除了只有一只手的某些专业棒球选手(例如Jim Abbott，www.jimabbott.net)，您其实不会看到棒球球员接完球后脱下手套用同一只手来投球。这是NoSQL和SQL的比喻；它们各自都针对特定任务和数据进行了优化，用两只手打球非常重要。

小结

在本章中，我们提供了一个总体框架，您可以用它来识别大数据。我们解释了可以从该框架获得的增强分析功能代表着一个转折点。我们使用数量、种类、速度和真实性这四个术语，让您用一种容易记住的方式来理解和分类大数据机会。大数据时代的重点是通过专用的大数据平台加强分析，该平台引入了全新的方法和技术，并让它们与同样重要的且目前已经在运行的传统解决方案协同工作。因此，大数据讨论从未以“让我们迁离目前的仓库技术”开始。这些方法彼此补充，就像协调多种技能，表现非常出色的运动员一样，IBM大数据平台也会推动各种让您获得成功和实现业务差异化的大数据措施。

2 将大数应用到业务问题: 使用情况示例

本章的标题描述了我们要介绍的内容: 如何应用大数据来帮助您解决业务问题。我们将使用多个精选的用例，以及对配合大数据使用的机制进行评论，从而阐述大数据提高业务绩效和解决以前难以逾越的(或极大地简化了困难的)问题的能力。我们将涉及如何帮助客户开发新的应用程序，可以使用哪些潜在方法来解决以前困难的挑战。我们还将讨论全新的做事方式如何从根本上改变解决问题的方法。

何时考虑大数据解决方案

您是否受限于目前的平台或环境，因为您无法处理自己想处理的数据量？您是否希望在分析范式中包含新的数据源，但却无法实现，因为在不牺牲数据的保真度或丰富性的情况下，它无法融入模式所定义的行和列？您是否需要尽快摄取数据，并需要使用一个按需模式(schema-on-demand)的范式？您是否被迫使用一个写时模式(schema-on-write)的方法(在加载数据之前，必须创建模式)，但需要快速摄取数据，或者在一个发现流程中，您希望获得读时模式(schema-on-read)方法的成本优势(数据只是被复制到文件存储，并不需要进行特殊的转换)，直到您知道自己已经获得可随时进行分析的东西？数据是否太快到达您的门槛，让目前的分析平台来不及处理？如果您对上述任意一个问题回答“是”，就需要考虑大数据解决方案。

在第1章中，我们概述了可以帮助您找到大数据的特征(速度、数量、真实性和种类)。我们还指出，您可能听说过的大部分全新大数据解决方案更有可能是补充性的，而不是替代您目前的分析平台(如可信的数据仓库)。总之，这种补充性的方法可以帮助您提高公司的大数据IQ。

让我们从一个很好的示例开始，看看我们如何利用大数据技术帮助解决业务问题。我们帮助一家金融服务行业(FSS)的大型公司了解为什么他们遇到了日益增加的减员和客户流失率。我们首先研究了来自其客户服务中心的整整一个月的电子邮件(约4千万条消息)。在本例中，从任何角度来看，所存储的数据量都不大，但我们仍然使用了IBM InfoSphere BigInsights (BigInsights)，这是IBM非分支的Hadoop发行版，因为执行分析需要强大的计算能力。还需要另一种方法快速、准确地咀嚼电子邮件。此外，我们不知道会发生什么，所以我们希望在规模、信息类型，以及我们可用于执行分析的方法方面具有最高的灵活性。我们还不知道我们会发现什么，并且我们需要“进入”其他数据源来获取洞察的灵活性；在本例中指Web

浏览行为、结构化的会计数据，以及帐户的绩效历史记录。这种灵活性往往是解决问题和推进项目的关键。我们以这个示例开始，因为它突出了一些常见的并且反复出现的用例模式。特别是，该用例提供了以下模式的示例:

- 需要从小规模历史记录(月)改变为数年
- 包括混合的信息类型，在本例中是结构化的帐户历史记录与电子邮件
- 跨系统的工作流，需要用不同于最初数据准备时的一种脚本语言来完成特定文件的导出格式化工作
- 利用计算密集型的NLP(自然语言处理)和机器学习技术。

当然，在此用例中完成的工作所带给我们的挑战似乎每次都会出现，最终我们建议您使用一种大数据解决方案，具体如下:

- 大数据解决方案不仅非常适合于分析原始的结构化数据，也非常适合于分析多种来源的半结构化和非结构化数据，我们的洞察来自所有这些来源的交集。
- 如果您不满意算法或模型的有效性；需要对所有或大部分数据进行分析；数据采样不具备同等有效性，大数据解决方案都是理想的。如果您不满足于只是对数据进行采样，还想看看每一次交互，因为您认为它会带来竞争优势(这就是我们在第1章谈到的总人口分析的主题)，大数据解决方案也是有帮助的。
- 如果有关数据的业务措施不是预先确定的，大数据解决方案对于迭代和探索性分析很理想，我们特意采用了一个迭代/敏捷的学习方法。
- 如果您无法完全确定调查将带您去哪里，并且希望获得计算、存储和将要进行的分析类型(随着我们增加更多的资源和新方法，这一切都会变得有用)等方面的弹性，大数据解决方案是理想的。

此外很重要的并且真正重要的是: 注意我们扩充了现有的分析投资的能力。在大多数情况下，淘汰和更换是不恰当的；相反，与我们合作的客户希望增加其现有投资的收益。在此客户案例中，我们使用了一个大数据解决方案，以利用不适合其传统数据仓库环境的数据，以及我们从仓库所知道的信息。

大数据技术也非常适合于应对本身无法用传统关系型数据库方法处理的信息挑战。重要的是要明白，传统的数据库技术是至关重要和必不可少的，是整体分析解决方案的一部分。事实上，在与大数据平台结合使用时，它们变得更加重要。我们常说: ″大数据不只是Hadoop。″ 话虽如此，有些问题无法通过传统数据库解决，至少在开始时做不到。有些数据是我们无法肯定是否希望保留在仓库中的，因为我们不知道其价值是否丰富——它是非结构化的，它的数量过于庞大，也许其保质期太短，或者要知道如何规范化其结构可能言之尚早。

在开始之前: 大数据、拼图和洞察

在深入讨论用例之前，让我们谈谈解决问题的大数据方法如何比常规方法更有效。在阅读本节之前，我们会要求您考虑以下两个问题: "在处理大数据时，寻找模式是更容易还是更困难？" 和 "随着数据量和速度的增加，发现异常值是更容易还是更难？" 在阅读本节之前先考虑一下答案，然后在阅读完本节后回答这些问题。

首次打开2000块的拼图盒，并将其中的拼图块全部拿到桌上时，所有拼图块都混在一起。通过应用一些基本的分组，您可以很轻松地找到直边和角部的拼图块，组成拼图的外框。完成拼图的外框后，仍然有大约1,900块要填充进去；和数据一样，需要使用分析来完成整个图片。

您可以将这个简单的模式匹配视为传统的报告，它侧重于可预见的、形状整齐的数据，可以轻松地将它们组合在一起并进行排序。如果打开拼图盒时，所有的边界形状都包装在一个单独的袋子里，而所有黄色图块也都放在专用的袋子中，诸如此类，那么完成这个拼图不会很难。但是，拼图并不是这样包装的，数据也不是这样到达您的门槛的。

让我们假设此时您已经在拼图上花了几个小时。外框已经完成，中间也正在成形。您可以看到部分图片。拼图的哪种状态包含更多数据？在将它从盒子中倒出来，变成一堆图块的时候，还是在您可以看到一些(但不是全部)图片的时候？

我们猜想，很多人最初会觉得两者都含有相同数量的数据——毕竟，拼图块的总数并没有改变。但数据并不只是有多少行，或有多少PB信息；它与上下文和理解有关。实际上，随着拼图慢慢成形，我们能更好地理解问题领域。我们可以进行推理，发现关系，这可为我们提供丰富得多的信息。也就是说，即使我们有相同数量的拼图，但我们现在要管理更多的数据(和元数据)。让我们来解释一下: 就像数据组合在一起可以产生更多洞察一样肯定的是，必须对新的关系和模式进行管理(这不一定和拼图有关系)。因此，在拼图完成35%时，比刚刚将它倒出盒子时有更多的信息，这是很客观的。虽然大家最初都认为处理更多的数据会更难、更具挑战性，但随着模式的发现，更多的数据会让问题更容易解决。

现在，假设有人在您不知情的情况下扔进了来自另一个拼图的一些图块。(本实验来自IBM的实体分析首席科学家Jeff Jonas；他的研究主体是毫无戒备之心的家人和朋友。要提防Jeff会在晚饭后搞乱一个拼图呀。)面对其中有5%无法拼起来的一堆拼图块，什么时候会比较容易哪些拼图块不属于这个拼图呢？在最初的图块堆中，还是在某些图块已经被放好的时候？当然，已放好的图块有

助于说明模式，让我们更容易找到那些外来者，并通过消除这些外来者更轻松地解决剩下的拼图。有更多的数据让我们更容易找到外来者。仔细想想:拼图越完整，就越容易识别模式，越容易找出外来者(另一个拼图的图块)。看看组织如何可以充分利用更多的数据？

现在让我们假设，我们要求您解决世界上最大的拼图，它有32,256块(由Ravensburger AG在2010年制作)。两个人需要多久才能完成？如果我们增加两个人(处理器)，并向他们安排互补但不同的任务，如寻找其中包含脸部元素的作品，这又需要多久才能完成呢？如果我们再增加另外两个人，他们唯一的任务就看看拼图的另一个独立部分，以此类推？您应该明白了吧，这正是我们使用机器向外扩展工作的方式，而不是依靠数量有限的处理器。现在，我们希望您开始明白为什么新的大数据技术对于完成拼图如此重要，拼图就相当于下一个最佳报价、下一个最佳行动、治愈一种疾病等。我们无法开始揣测有多少块数据属于大数据时代拼图的一部分，所以需要向外扩展、机器学习和大规模分析系统来确定其边缘，对数据块进行排序和分组，并发现模式。

您可能想知道在本节开始时我们要您考虑的问题的答案：现在我们的答案很可能与您的答案是一致的——更快、更轻松。利用大型数据集的解决方案往往比受限于较小数据集的解决方案更有帮助。

大数据用例：大数据部署的模式

我们讨论过在本章中要介绍哪些用例，我们试图找出哪些行业没有适合大数据的用例。当有人脱口而出“DJ行业”时，我们认为找到了优胜者，但之后，在BigDataUniversity.com有一个学生将他的整个音乐收藏放进BigInsights，随后对他的音乐文件构建了一些很酷的分析，包括面向音乐基因组学的应用程序。讲了这么多如何使用该方法缩小范围，这导致我们选择的用例可以帮助您挖掘常见并且在各行业都会出现的模式。例如，接触新的数据类型并在数据流离开装配有传感器的采集单元时应用分析，这样做可以获得对交通系统、网络、新生儿健康、贷款账项估值等的洞察。

此处详述的使用模式涉及BigInsights、IBM InfoSphere Streams (Streams)，以及新时代的大数据引擎与传统仓库的组合，如IBM PureData System for Analytics(前身为Netezza)或IBM PureData System for Operational Analytics平台。在我们介绍的所有用例中都有一个共同的趋势：它们都涉及到利用大数据平台，以更实用(并且现在终于成为可能)的全新方式来做事情。

您花了钱实现物联化——现在充分利用它吧！

在第1章中，我们为您介绍了各种指标和示例，说明我们的世界物联化程度有多高：从

桥梁，到路轨、牲畜、会议徽章、跑步鞋等，这个世界几乎可以在任何地点、任何时间从任何东西处收集某种数据。

以一个典型的石油钻井平台为例子，它的主板上可能有20,000至40,000个传感器。所有这些传感器都以流式方式传输有关石油钻井的状况、作业质量等方面的数据。并不是每个传感器都全天候主动广播数据，但也有一些传感器每秒要报告多次。现在猜一猜主动利用这些传感器的比例有多少。如果您认为在10%(甚至5%)的范围内，您要么很会猜，要么已经了解跨越行业和用例反复出现的大数据主题: 客户在其决策制定过程中并没有使用为他们提供的所有数据。当然，涉及到能源数据(或该主题的任何数据)的收集率时，它确实引出了一个问题: ″如果您费尽力气物联化用户、设备或钻机，从理论上讲您是有意这样做的，那么，为什么您不捕获并充分利用所收集的信息呢？″

在这种使用模式中，重点是对所获得的静止数据应用分析，并将分析结果应用于运动中的数据，从而更好地理解该领域。The University of Ontario Institute of Technology(UOIT)首席研究员(Carolyn McGregor博士)与多伦多的The Hospital for Sick Children进行合作，找到一种更好的方法来预测可影响新生婴儿某种特定院内传播疾病的发病情况。您可以想象这些脆弱的婴儿被连接到持续收集数据的机器会是什么样。有些医院记录每小时或每半小时的读数，并在72小时左右后丢弃；从而没有能力发现静止数据中的趋势，并以更精细的水平将分析应用到运动中的数据。UOIT的首席研究员Carolyn McGregor博士利用了IBM的Streams技术创建一个运动数据分析平台，每秒分析超过1,000条独特的医疗诊断信息。想象一下，120名婴儿的传感器数据量，这就相当于每秒分析12万条消息，每天分析1.788亿条消息！您可以在Web上找到有关这个美妙成功故事的详细信息(搜索″IBM data baby″)。现在，想想将这种方法扩大到门诊患者，他们挂上一个传感器，可以随时了解他们的日常活动，或监控有可能进入慢性疾病状态的人。简单来说，大数据有可能改变游戏规则，也有可能成为生命的救星。

看看这个世界有多少东西是物联化的，因此传感器数据是惊人的(电网、石油钻井、交通流量、收费路线等)，这意味着它们的意图是收集数据: 大数据现在让您可以对数据进行一些处理。

IT对IT: 数据中心、机器数据和日志分析

基于Hadoop的日志分析已成为一种常见用例，但是，这并不意味着其部署如预期般广泛。日志分析实际上IBM与多家公司合作后建立的一个模式，最初在FSS中建立。之后，我们看到这种用例出现在各个行业；因此，我们将这个模式称为IT for IT(IT对IT)。

通过大数据得到充实的“IT对IT”可以帮助客户更好地了解其系统如何运行，各部分何时以及如何被拆分。例如，某金融公司将搞清楚应用程序如何运行的传统方式亲切地称为“打鼹鼠”。当在其严重依赖于SOA的环境中发生了问题，总是很难确定到底发生了什么事，因为一个给定交易的处理涉及超过20个系统。(我们都看过这部电影，每个人都在作战室跑来跑去，嚷嚷着“不是我做的！”那部电影里还有一个场景，每个人的手指都指向……您！)

我们使用该模式帮助的其中一个客户最终获得了每天分析大约1TB日志数据的能力，其时延不到5分钟(这个用例同样适用于更大或更小的日志生成速率)。如今，该客户能够破译在其整个IT体系的每一个交易中到底发生了什么。如果他们的客户从其移动银行或网上银行站点发起某个交易失败了，他们可以说出发生失败的确切位置，以及是什么组件造成了这个问题。正如您可以想像的，这拉平了解决时间指标。

此客户可以利用他们从静止数据中获取的洞察，并充分利用Streams对其网络的状况进行的实时调查。例如，如果在Hibernate(休眠)层中内存堆的不断消耗与应用服务器中的堆栈溢出密切相关，能够在这个问题的起源发现它就可以保住交易或防止网络中断。例如，我们其中一个电信客户使用Streams实时分析机器数据，以发现行为异常的、损害网络的移动计算应用程序，从而让有问题的应用程序可以被终止。

有时候，我们喜欢将因运营IT解决方案而生成的所有日志和跟踪数据称为数据废气。企业有大量的数据废气，像任何污染物一样，它就会被丢弃，因为它被视为废物。日志数据往往与高存储成本紧密联系在一起。然而，日志中包含了大量潜在的洞察，不仅仅关于您的服务器现在发生了什么事，还关于将要发生什么事。

仔细考虑一下，组织如何编制基础架构预算。您愿意认为它是基于事实的，但这是真的吗？您是否有一个清晰的跨端到端平台的系统利用率视图，其中包括趋势？您是否了解在各团体和部门间的季节性因素和其他活动的影响？我们能够向我们的其中一个客户证明，他们计划采购新服务器所依据的峰值负载量，其实可以利用现有的空闲系统来处理。我们帮助他节省了数百万美元，结果在第一年就获得了三位数的投资回报率。

我们帮助另一个客户建立一个集中式的日志信息交流中心，而不是让每个子小组部署自己的解决方案。该客户在为期两周的窗口中存储日志，然后删除它们以规避存储成本。依赖于这些系统的部门和应用程序开发团队知道日志将被删除，因此会获得这些日志并将其放在昂贵的SAN上。没有企业范围的保留策略，各部门使用日志的方式也没有一致性。当我们思考这种情况时，是很讽刺的。这些日志被删除，是因为它们产生了存储费用，但它们最后以一式三份(或更多)的形式

被放在没有计划好删除机制的昂贵SAN上。该客户使用BigInsights创建一个集中式日志信息交流中心，并实现了超过一百万美元的节省。具体来说，他们在BigInsights存储上的套利抵销了SAN的成本，并实施了9个月的滚动保留策略；在这段时间之后，日志的保留价值不大。但是，他们并没有就此停止。现在，日志都在一个地方，并且保留一段合理的时间，让他们能够确定整体的趋势和问题；总之，他们能够将各个点连起来了。

但该用例并不仅仅要说明检测问题。该客户目前正在整理大量知识，让他们能更好地预测失败，并理解失败之间的相互作用。他们的服务部门可以针对具体问题生成最佳实践补救措施，或调优基础架构来消除问题。这就是可发现的预防性维护。我们的一些大型保险公司和零售客户需要知道“失败的前兆是什么？”或“这些系统之间有何关系？”等问题的答案。这些是传统的监控无法回答的问题类型；大数据解决方案最终让您有机会对手头上各种问题实现更好的新洞察。

什么、为什么和谁？社交媒体分析

也许人们谈论得最多的大数据使用模式涉及社交媒体和客户情绪分析——它也可能被过分夸大和误解。虽然大数据可以帮助您找出客户对您的品牌(或竞争对手的品牌)说了些什么，但我们认为需要拓展目前的工作重点。

社交媒体分析是一个相当热门的话题，但我们已经开始看到“买方疲劳”，因为目前的实践没有支持有关该用例的不断炒作。简单地说，人们说或想什么，与为什么他们会这样说和这样想，这之间有很大的区别。您的社交媒体大数据分析项目试图回答“什么”和“为什么”，提供您所追求的分析收益。

最近，我们能够识别出专门针对我们所帮助的一家金融公司的一些负面议论。我们想找出一开始为什么会出现负面议论，为什么人们会进入使用技术来传播这种消极情绪的状态？更重要的是，这是否会影响销售，可以做些什么来改善这种情况，如果有一种特定的响应可以改善情况，我们又如何知道？我们将这称为闭环分析。如果您只是听见人们说什么，但不具备对信息进行分析并作出相应反应的能力，您的行动很大程度上仍然是盲目的。

了解为什么人们这样评论您的组织，这包括与之相关的一切:宣传、产品组合、价格变化、政策变化、市场营销、企业社会责任，以及最初促成消费者意见的一系列其他活动。还有一个少数企业正在谈论的基本要求:您需要在相同的分析管道中结合外部和内部的信息流。这就是您开始获得真正的洞察和实现提升的地方；事实证明，只有很少的(如果有的话)外部服务提供社交媒体产品来做到这一点，因为他们不能处理内部的结构化数据来源，或者他们缺乏将二者结合起来的分析——IBM大数据平台可以做到这一点。

了解客户情绪

某些行业或产品的客户并不忠诚，具有非常高的流失率。我们曾与一家电信公司合作，该公司经历了每年50%以上的客户流失。他们希望在计费周期中更早地识别出哪类客户是最脆弱的，从而提高保留率。对于这个客户，即使其用户转换到更稳定计划的比率有很小的提高，也可让他们其中一个产品的收入流增加一倍。很简单，他们并不需要一个本垒打才会取得更大的成功；他们只是需要一个二垒安打。当然，通常说起来容易，做起来难，但是，机会对于客户而言转瞬即逝。处理这种级别的流失率和大量的数据时，能够实现数据发现、捕获、响应和互动就是一个不小的挑战。

实现这一点的关键是，能够检测到忠诚度在下降，并在与客户的下一次联系之前将它融入到您的客户协议和下一个最佳行动模型中。这样做不能算实时处理，而是我们所说的客户时间。客户时间只是一个概念，在下一次客户互动之前能够处理所需的一切，让它在客户的眼中是无缝的。然而，随着智能手机的普及，客户时间越来越接近于实时，因为总是有机会让他们了解情况，例如发送电子邮件、文本信息或优惠。

除了在客户时间内操作，此用例提供了另一个示例来说明捕获所有可用信息的价值，以观察有助于建立上下文的事件。在您捕获一个方面的通信内容后，就可以继续到下一个方面，并开始将它关联到从电子邮件到社交媒体及我们已经在本章讨论过的其他东西；您甚至可以将它关联到后端办公服务质量报告，根据您的后端系统，看看是否有人致电并表示对您的不满。如果能够识别出那些表明您的系统速度较慢或行为异常的模式，而且这恰好是某个特定用户致电取消服务而没有明确提到的原因，您就可以同客户所说的内容建立关联。事实上，我们其中一个FSS客户的目标之一是在他们跟您谈话之前，就深入了解为什么您要打电话，让他们能够预先解决您的问题，并将修补计划告知您，而不是与您谈论问题！

我们相信，您可能对以下场景深有体会：您在电话中断两次后致电服务提供商的客户服务部门。您必须重新进行身份验证，然后向另一位座席代表重复整个故事，因为当您的电话出现中断问题时，原来的座席代表还没有完成记录。现在想象一下，再次打电话过去并听见座席代表说，“很抱歉，我们的电话有些问题，并注意到您已经被中断两次。我很抱歉，我必须对您重新进行身份验证，但我理解为什么您打电话过来，以下是我们可以做的，以帮助解决这个问题……”我们愿意打赌，这种回答可能会超出您的期望。

另一个示例：有多少次您致电Internet服务提供商投诉您的高速Internet服务，而得到的印象却只是客户服务代表(CSR)的工作是让您觉得自己并不重要，并尽快让您挂断电

话？从业务的角度来看，您必须知道提供商是否真的捕获了服务问题。也许处理电话的座席代表填写一个表单，概述基本的服务投诉，但它会与显示系统运行情况的时间点定量报告有什么关联吗？如果您有一个大数据平台，您可以获得洞察和并在投诉出现之前预测问题。在Internet服务的情况中，服务质量(QoS)的测量很简单——该供应商的技术支持部门是否分析Internet协议的详细记录(IPDR)，从而了解网络的状况，然后每当城市的某一部分存在QoS问题时就会提醒客户服务部门？

诚然，之前的场景对于今天的呼叫中心来说是非常先进的，但在大数据世界中，它还有成长的空间，如音调变化或文本分析，以识别愤怒的情绪(“我已经第三次打电话了！”)或进行实时关联，以识别情绪的趋势，确定如何在呼叫中心将趋势与其余的业务操作相关联。如果您需要解释这是第三次不得不致电来投诉同一个问题，即使您选择使用不同的渠道，与供应商之间的所有交互不是都应该反映出这一点吗？

了解客户情绪是一个非常有趣的大数据用例，因为可将其应用到目前有可能出现的所有数据上(对运动或静止的数据使用分析)，也可以应用到新出现的功能。您可以使用其中一个IBM大数据平台静止分析引擎(如BigInsights)发现和构建自己的模型，并获得业务洞察。然后可以选择继续使用静止分析，用低得多的延迟获得电话交互，或者构建这些模型，然后将它们重新提升到业务前沿，使用Streams研究和分析电话内容并尽快转换它们，以近乎实时地获得洞察。这将业务从预测(我们认为，如果……客户会离开)转换为现报(我们知道这个客户将会离开，因为……)。Streams分析的结果流回BigInsights，创建一个闭环的反馈机制，因为BigInsights可以遍历各种结果，以改进模型。

社交媒体技术让您随心所欲

在前面的章节中，我们谈到了社交媒体分析，说起了很多传言和疲劳——它可能被过于夸大了。我们认为，我们会帮助您扩大适用此技术的范围；事实上，如果您拥有从非结构化文本提取结构和理解内容的技术，就能想象无限多种使用模式。

用于社交媒体的文本分析技术同样可用来做一些不同的事情。例如，一个客户想调查对其知识产权的盗版传播和版权侵犯(它是基于视频的)。这是一个非常有趣的领域，因为在社交媒体上常常有对盗版类型的讨论，这个客户能够建立模式来识别在全球各种微博站点中发生的这种对话。这最终将他们带到包含非法发布了属于客户的受版权保护材料的网站。他们如何做到这一点呢？他们开发了字典和模式来识别在“被盗”材料中的运动队名称，“下载”有关免费内容的诱人链接的语言，对缩短的URL(tinyURL)进行URL解析，获得实际位置等。因此，虽然他们并没

有分析谁说了什么，但的确通过基于非结构化文本的数据创建了关于问题领域的结构。同样的技术、同样的工具集、同样的方法，只是不同的用例。

我们介绍了一个供应商在限制其知识产权盗版方面的行动，您可以看到大数据分析平台(如IBM提供的产品)的多种应用方式。例如，一个“监察”客户筛选有关国家议会的出席和投票记录的无数网页，产生很强的相关性，并将注意力集中在特定的一位代表在其当选位置上的行为(在本例中，由于该代表所缺少的票数，给人留下了深刻印象)。

另一个客户使用相同的概念构建一个有关情绪的废话字典，并应用它来创建对日志文件的理解(我们会在本章后面讨论这种使用模式)。例如，一个IP地址本身具有一定的结构(IPv4或IPv6)，监控代理有自己的名称(如Nagios)，依此类推。应用程序服务器和数据库服务器上的监控代理名称之间是否有相关性？

最后，某投资公司搜索IT公司基于文本(HTML)的公开收入报告，以查找披露内容中有关业务状况的“秘密信息”或迹象，或在评论中有关其未来业绩等的指导因素。您可以想象，比较不同公司(如HP和IBM)的服务收入流可能是非常有用的，但每个公司很可能对这些部门有不同的命名，例如，HP称之为“惠普企业服务”，而IBM称之为“IBM全球服务”。手工组装大规模市场参与者的这些信息是一项艰巨的任务，但如果构建文本提取器来定义如何识别公司的名称和服务部门的名称，然后也许使用Nutch(一个基于Apache Lucene的开源Web搜索引擎)在Web上进行爬网，下载与感兴趣的每家公司的该部门有关的评论，这就会成为一个轻松得多的任务。

客户状态: 或者，不要在我生气的时候试图向我推销产品

研究似乎表明，无论是由于外包、紧迫的解决时间指标、经济因素还是经费削减等，客户服务正变得越来越差。许多人发现很难想起他们最后一次良好的客户体验是什么时候出现的。服务质量往往取决于个人，而不是业务。(我们已经有这样的经验: 因同样的信息或问题致电产生不同的结果，这取决于CSR的态度和培训)。我们中的许多人已经习惯于以平庸的客户体验为规则。但是，并不一定要这样，消费者的声音也越来越大，他们不想因为与他们做生意的公司在经济困难时期削减服务成本来提高其盈利能力而受罚，所以有些事情必须改变。我们一直忙于帮助企业处理这件事，好消息是，在我们的经验中，企业希望提供良好的服务，而大数据技术在帮助改善服务的同时也可降低成本。

关键是通过利用所有可用信息，重新构想整个流程如何工作：帮助业务在正确的时间，在正确的上下文中做正确的事。当然，这不是一个新目标，但大数据极大地提高了真正做到这一点，并把它做好的能力。

大多数客户参与数据都被忽略了。如果有的话，也只是从呼叫中心交互中捕获了极少的上下文。通过公司网站的点击流数据也是如此。这难道不是一种沟通形式？在本例中，客户在说您所做的足以引起他们的兴趣，并希望了解有关您的公司或产品的更多信息。通常在汇总级别上使用这种类型的数据，我们将客户的行为视作一个整体；例如，您的客户在查看什么产品，什么被添加到购物车中，什么样的购物车被放弃。为什么不改为在更加个人的级别上使用这些数据，以发现客户实际上在做什么？例如，购物车是否一直在订单流程的运费计算阶段被放弃，或者经过搜索之后似乎并没有产生结果？通常不会保存和分析这种类型的数据粒度，因为它需要太多存储，也许认为它的保质期过短，不值得投资，或其处理速度不够快，无法在个人客户级别使用。再举一个例子，试想您最后一次向服务提供者发送电子邮件，并且它真正改变了他们与您的交互方式。不应该吗？您敢打赌它应该这样！在大数据时代，那些迫使人们进行折衷，从而降低服务水平的存储和处理问题可以开始得到解决。

考虑在一个场景中，您可以结合仓库对具有以下通信事件的客户进行了解：您最终会对我们所说的客户状态有更加丰富、明智、及时的了解。本书的一位作者最近经历了一个很好的客户状态不匹配示例。

他和一家公司之间发生了一个问题，他的信用卡计费出现了错误，但银行告诉他，他必须找供应商解决问题。银行不会停止计费，因为这是一笔预授权费用。在几天的过程中，他发送了电子邮件，然后聊了聊，然后打电话到银行，挫折感每次都在不断提高。这种解决方法没有结果之后，他走进一家地方支行，对方即刻试图向他推销一张新卡的功能。这显然不是向他进行推销的适当时间(他的客户状态不对)，但可怜的CSR完全不知道这是做这件事的错误时间。这种银行体验远远称不上独特(并且我们的作者希望您知道，他们通常处理得相当不错)，但这是一个事实：大多数企业没有捡起客户发给他们的信号。

欺诈检测：“谁在凌晨4点购买订婚戒指？”

您最后一次在凌晨4点买订婚戒指是什么时候？(我们不包括在拉斯维加斯的任何采购)。不是很常见，对吧？这是异常值概念的一个很好的示例，它是寻找和试图预测各种诈骗行动的关键。您总是要搜索异常值。试想一下智能手机的详细清单，您肯定不会查看数百个电话来试图弄清楚如何挥霍掉2,000分钟；但我们敢打赌，如果其他通话分钟数都是个位数，您就能发现那次70分钟的通话。欺诈和风险属于跨行业的大数据用例，它并不仅限于FSS。正如我们将看到的，大数据平台非常适合于寻找异常值。更重要的是，高度动态的环境通常具有周期性的诈骗模式，在数小时、数天或数周中反复出现。如果

无法以低延时提供用于识别新欺诈检测模型的数据，当您发现这些新模式时就为时已晚了，并且它已经造成了一定的伤害。

欺诈检测的某些挑战就是只使用传统的技术。所有大数据模式中最常见的主题仅限于可以存储什么(数量和类型)，以及提供什么计算资源来处理您的意图。换句话说，预测欺诈的模型往往要么过分聚焦在计算限制和错过的东西，要么就无法达到预期(或应该)的精细程度，因为模型的维度肯定会有人为的约束。毕竟，如果不存储数据，分析这些异常值的属性，就很难找到异常值。这就是说，只有您具有梳理数据并找到埋在噪声中的信号所需要的计算能力时，更多的数据和属性才是有用的。以足够快的速度加载和处理数据，捕获快速移动的事件，这也是至关重要的。

传统的诈骗案件涉及使用样品和模型来识别表现出一定特征的客户。虽然这种方法是可行的，但它的问题是(这是在很多这些用例中都会看到的趋势)，您分析一个市场细分类别，并且不是在单独的事务或个人级别进行分析。基于细分类别进行预测是好的，但根据个人的实际资料作出决定并关联他们的交易显然是更好的。为此，您需要处理的数据集比传统方法可以处理的数据集更大。我们估计，在可用的信息中，只有不到50%(通常比这个数字还要少得多)对欺诈建模可能有用的信息被实际使用。您可能会认为，该解决方案是将另外50%的数据加载到传统的分析仓库。这是不实际的，其原因似乎出现在多数大数据使用模式中，即: 数据不适合；它会包含仓库无法有效地利用的数据类型；它很可能会要求破坏性的架构变更；以及它很可能将现有工作负载的速度减慢到像蜗牛一样。

如果将剩下的数据都填进现有的仓库是不可行的，那怎样才可行呢？我们认为，IBM大数据平台的核心引擎(BigInsights、Streams和基于分析的IBM PureData Systems)为您提供了灵活性和敏捷性，将您的欺诈模型带到一个全新的水平。BigInsights解决了我们在上一段中概述的问题，因为它会扩展到几乎任何数量，并处理所要求的任何数据类型。因为它没有施加一个写时模式，您将在如何组织数据时拥有最大的灵活性，并且工作不会影响现有的工作负载和其他系统。最后，BigInsights是高度可扩展的；您可以从小规模开始部署，并以非常经济高效的方式增长(如果我们说您的CIO肯定会喜欢这一部分，请相信我们)。

现在，您拥有了BigInsights，可为所有可用数据提供一个富有弹性且经济高效的存储库，您如何去找出那些异常值呢？

我们在这里尽量保持简单，因为这个话题本身可以写满一整本书，但第一步是将全部有用的数据加载到BigInsights。请注意，我们没有通过数据类型来限定数据的资格。第二，构建一个基本的客户档案，包含尽可能详细和尽可能多的行为维度。第三，开始

构建模型，定义“正常”是什么模样，然后开始搞清楚异常值的模样，以及它需要有多“异常”才值得将它进行标记。在第一次迭代中，我们预期某些异常值是不正确的，但以迭代和经验为基础的学习是我们所追求的。当然，回溯测试是这一过程的一部分，但对当前数据流进行测试也同样属于该过程，您可以通过将模型推送到Streams来做到这一点(Predictive Model Markup Language是做到这一点的其中一种方式)，这样您就可以对一个实时数据流进行评分，以加快学习过程。随着发现和验证各种模型，您会将它们推广到日常的高性能分析平台，这就会将传统的数据仓库技术带进画面。寻找异常值将提供有关“正常”客户在做什么的额外洞察，而客户细分模型没有因此而改善是很少见的情况；当然，在改善发生之后，您也需要释放您的团队，让他们开始使用这些数据。

我们在本章开始时说过，有数百种使用模式，但我们在本章中对它们进行一一介绍。事实上，欺诈检测有很高的适用性。想想在医疗保健市场的欺诈行为(健康保险欺诈、药物欺诈、医疗欺诈等)，并且走在保险公司和政府的诈骗计划(原告和供应商)前面。这是一个巨大的机会；美国联邦调查局(FBI)估计，医疗保健欺诈一年就花掉美国纳税人600多亿美元。想想欺诈性的在线产品或机票销售、汇款、银行卡刷卡等。

流动资金和风险: 从汇总到个人

在许多行业中的风险建模和管理都有改进的余地，这个事实可能是令人震惊的。涉及到在整章中所讨论的大数据使用模式时，风险建模带来了对反复出现的问题的关注: “您在建模中使用了多少数据？”

和“迭代和刷新这些模型需要多长时间？”目前的建模方法受到了系统的约束，它们决定了架构并将获得最佳效果的路径排除在外。这并不是说公司没有认识到在未被发现的数据中可提供很多潜在的洞察，但他们目前的系统并不总是支持他们。分析仓库的大小增加两倍、三倍或四倍是行不通的，因为它们通常已经拥挤不堪或已被充分利用了。资本约束和监管要求正迫使在这个领域中采用有趣的新方法，我们会在本节中与您分享一些新方法。

以一个跨国公司客户为例，需要从按业务线组织的管道转变为个人层面的风险管理。他们目前的数据是一个月时间的信贷风险快照；肯定是有用的数据，但以他们目前的手工流程来处理是非常缓慢而且昂贵的。我们提出了一个新的架构，其中包括BigInsights、IBM Information Server和Netezza(IBM PureData Systems for Analytics当时的名称)。我们使用了他们现有的IBM Information Server Data Stage平台来转换其原始数据，并将充实后的数据加载到HBase(在BigInsights产品中包括的

面向列的数据存储)。HBase让您能够以键/值对的形式保存和表示数据。以时间序列写出客户信贷风险，在提供新产品时该序列被扩展。使用这种方法，BigInsights能够在一个大小和性能均具备弹性且成本明显降低的环境中维护客户信用状况的当前表示。Netezza分析环境可以在需要时请求来自BigInsights的数据，以充实模型，而无需担心客户如何改变获得并存储其数据的流程。

小结

我们不可能在这么短的篇幅内评判这个主题。但底线在于: 仅仅是有更多数据并不能为您提供更深入的分析洞察，但在整体上实际加速了数据分析过程，您只需要有一个平台可以帮助您能处理更多数量、种类、真实性和速度的数据。IBM大数据

平台就是这样一个平台，因为它为运动中的大数据和静止的大数据提供了搜索、开发、治理和分析服务。

我们希望，我们在本章所包括的少数大数据用例可以帮助您认识到，将更多高保真数据添加到分析过程并将数据的覆盖范围扩大到包括传统和非传统来源，会让您获得一个高度自适应的分析模型。

3 提高您的大数据IQ: IBM大数据平台

如今的组织都了解分析的价值。在第1章中，我们谈到了分析(作为在各个市场细分中识别最终领导者的一种手段)所造成的影响。我们谈论了IBM/MIT Sloan的联合研究The New Intelligent Enterprise(新的智能企业)，它清楚表述了市场的领导地位和企业的分析IQ(大数据可以提高它)之间有着很强的正相关性。事实上，客观来说，大部分围绕大数据的关注都与“提升”其分析IQ，使自己在市场中与众不同，从而获得既得利益的公司有关。正因如此，许多组织现在都认为，从数据中获取深刻的分析洞察已成为在全球经济中生存下去的当务之急。

新时代的分析

现在问问自己：“如果越来越多的组织采用分析，会发生什么情况？”他们是否会因为都在利用更多的洞察进行竞争而失去了自己的竞争优势？例如，在十年前，Oakland Athletics(奥克兰运动家棒球队)采用分析和数据驱动的方法进行球员评估，他们用这种方法获得了职业棒球大联盟(MLB)中其他球队所没有的优势。事实上，该队在其同行中是领导者，展示出卓越的工资表现比和每座位净收入指标。他们使用分析让自己从业内同行中脱颖而出。事实上，他们的故事如此引人注目，引起了他人非常大的兴趣，以致于畅销书和Brad Pitt主演的电影大片Moneyball都讲述了这个故事。最终，其他球队也追了上来。今天，MLB中所有球队都在使用复杂的分析。分析在MLB中的广泛采用抵销了Oakland A′s在这一领域曾经拥有的任何优势。

客观地说，我们现在正进入这样一个时代，大多数组织将分析视为一个“赌注”功能。让我们看一个具体的例子，某组织如何将分析的使用带到一个全新的水平。在许多以客户为中心的组织中，使用分析来衡量客户支持的有效性是一种常见的实践。这让他们能够监控客户满意度，提高保留率，并对支持成本进行管理。传统的方法包括分析与支持事件相关的所有数据(如电话持续时间和解决的速度)，然后识别改进的机会。它还需要执行调查和收集满意度指标。其中一个这样指标是净推荐指数，这是一个非常有效的工具，可以根据客户与技术支持人员之间的交互，衡量客户对公司或产品的看法。

虽然这种传统的方法可以有效地提高客户满意度，并降低客户流失率，但分析周期(支持电话和被推送到前线的实际流程改进之间的时间间隔)可能会很长。在此期间，其他客户可能有类似的较差支持体验，这可能会导致他们流失。组织实现差异化和竞争力的机

会，不仅与其业务核心的深度分析有关，也与分析周期有关。就像是一个巨大的雪球滚下山，分析对业务的影响最初是缓慢的，但在每一次转动中，潜在的影响都会变得越来越大。

考虑到这一点，真正的问题变成，“是否有可能实时地将构建于历史数据集之上的分析模型和流程应用到流式传输的数据上？”

我们其中一个客户目前正在这样做。他们有一个智能拦截代理，监控客户和客户支持代表(CSR)之间的所有电话对话。此代理监视对话，将情绪分析应用到该对话，并实时向CSR提供建议。例如，如果客户使用音调转折来提出一个问题，或使用嘲讽表示不满，自动代理能够立即检测到，并向CSR提供具体的指导。

建议可能是用不同的方式回答这个问题，将电话升级到一个新的级别，向客户提供特定的激励计划，或者只是更有礼貌一点。

通过实时地拦截、监测和分析此类电话，该客户支持代表能够立即采取补救行动来提高客户满意度，从而极大地提高支持活动的有效性。我们想强调，这些功能并不能取代传统的离线分析；它们只是纳入新的数据种类(在本例中是语音)并实时执行分析，从而对传统的离线分析进行补充。

分析型企业的主要考虑因素

分析方面的差异，是指使用数据驱动的洞察来完善组织的战略，并通过以前不可能实现的方式来进行分析。过去限制组织在何处以及如何运行分析的限制因素被消除。此外，一些当今最典型的“分析型企业”正在改变其分析部署模型，寻找新的竞争优势，并让自己与同行有所区别。

在本节中，我们将与您分享企业分析的主要宗旨：如果您想提高自己的大数据IQ，必须从正确的配料开始。我们将使用真实的客户示例来描述这些“配料”。

对大型数据集运行分析

从历史上看，对大型数据集执行分析一直是一个非常繁琐的过程。因此，组织求助于在可用数据的采样子集上运行他们的分析。虽然他们构建的模型和他们产生的预测还不错，但他们觉得，使用更多数据可改进他们的结果。他们认识到，采样过程有时可能会导致错误或有偏见的结论。

可以按一定比例对其整个数据集运行分析的组织，与那些没有这样做的组织相比，肯定是有优势的。太平洋西北国家实验室的智能电网示范项目就是一个很好的例子。该项目希望推动一个充满活力的全新智能电网行业和更为经济高效可靠电力供应，这两者都是可推动美国经济增长和国际竞争力的因素。他们计划收集大量的数据(具体是来自五个州

60,000户仪表客户的事件数据)，并对它们使用复杂的分析模型。他们希望使用IBM的大数据技术能够验证新的智能电网技术和商业模式。

快速地运行分析

运行分析是一个多步骤的过程。它涉及到数据探索、数据清理和转换、创建分析模型、部署这些模型并对它们评分、发布结果，然后改进模型。这也是一个迭代的过程。如果在运行分析查询时，底层的分析系统性能不佳，这会增加整体过程的时延。因此，如果有能力以极快的速度对大型数据集运行复杂的分析，就会具有明显的优势。

首先，它能够提高业务敏捷性，并极大地缩短整体决策制定所需的时间。我们的一个证券交易所客户(纽约泛欧交易所)将在2PB数据上运行深度交易分析的时间从26小时缩短为2分钟！他们运行的深度分析查询需要大量的数据访问和计算。这种性能提升不仅帮助他们更快地应对市场变化，还让他们能够提高其分析模型的复杂性。

其次，它提高了分析师的生产力。另一个客户，Catalina Marketing(大型零售营销服务提供商)能够将其分析师的生产力提高六倍。他们通过将其分析模型的平均评分时间从4.5小时缩短到60秒来做到这一点。因此，他们能够使用相同的人工运行更多分析模型，提供更深入的、变革性的业务洞察。暂停一会儿，考虑如果能够将评分的模型增加一百个，同时试图预测用户流失、风险、欺诈、气候变化或野火的烟雾分散率，这会有什么影响——这是大数据的机会。

实时运行分析

有能力在事件发生时实时运行分析，这对于组织如何响应各种变化有着深刻的影响。正如我们在第2章中所述，UOIT对运动中的数据采用分析，其目标是在长达24小时之前就预测出有可能危及生命的疾病是否会出现，这对于患者的治疗效果可能造成巨大的差异。他们通过分析来自各种监视器和生命体征指标的流数据做到这一点。我们的一个电信客户能够实时分析流式传输的网络流量数据，以检测各种瓶颈，并能够对网络应用预防性的维护工作。通过将分析移动到“更接近行动”，组织营造了巨大的差异化机会。

在种类更广泛的数据上运行分析

在本章的前面我们介绍过，一个客户通过实时分析来自支持对话的语音数据，实现了出色的客户支持效率。纳入新的数据种类(如语音、文本、视频和其他非结构化数据类型)与结构化关系数据源的能力，开辟了提高效率和差异化的可能性。现在，我们的一个零售客户关联了社交媒体数据与其数据仓库中的销售点数据。在推出一个新品牌之前，他们知道它会产生哪些议论，他们使用这些信息来预测按地区划分的产品销售情况，并确保

商品库存能满足这种情况的需求。他们正在对需要大量计算的库存水平和模型运行深度分析查询。

大数据平台宣言

为了支持分析型企业的关键考虑因素，重要的是有一个大数据平台的迫切需求清单——这引入我们的大数据平台宣言。传统方法的局限性已导致项目失败、昂贵的环境和不可扩展的部署。大数据平台必须能容纳所有数据，并且支持各种业务分析所需要的计算能力,为了实现这些目标，我们认为，任何大数据平台必须包括六个关键的迫切需求，如图3-1所示。

1.数据发现与探索

数据分析的过程以了解数据源开始，搞清楚在一个特定的源中提供什么数据，并了解其质量以及它和其他数据元素之间的关系。这个过程称为数据发现，它让数据科学家能够创建合适的分析模型和计算策略。传统的方法要求将数据物理地迁移到目标平台，然后才能够被发现。在大数据的情况下，这种方法过于昂贵，不切实际。

为了方便数据发现并释放驻留在大数据中的价值，该平台必须能够发现“已就绪”的数据。它必须能够支持索引、搜索以及不同大数据源的导航工作。它必须能够支持不同数据源的发现，如数据库、平面文件、内容管理系统——几乎是包含结构化、半结构化或非结构化数据的任何持久性数据存储。

图3-1大数据平台宣言: 迫切需求和底层技术

不要忘了，必须严格遵守和保存底层数据系统的安全配置文件。这些功能可让分析师和数据科学家受益，可以帮助他们在其分析应用程序中快速整合或发现新的数据源。

2. 极高的性能: 在更接近数据的地方运行分析

传统的架构分离了分析环境与数据环境。分析软件在它自己的基础架构上运行，并从后端数据仓库或其他系统中检索数据，以执行复杂的分析。这样做的理由是，对数据环境进行了优化，以便更快地访问数据，但不一定适合于高级的数学计算。因此，分析被视为一个独特的工作负载，必须在单独的基础架构中对其进行管理。这种架构的管理和运营成本很高，并且造成数据冗余，随着数据量的增加而性能下降。

未来的分析架构需要在同一平台上同时运行数据处理和复杂的分析。它需要提供PB级规模的性能吞吐量，方法是在平台内无缝地对整个数据集执行分析模型，无需复制或采样数据。它必须让数据科学家能够更快地对不同的模型进行迭代，加快发现和实验过程，产生“最适合”的收获。

3. 管理和分析非结构化数据

在很长一段时间内，我们一直根据数据的类型对其进行分类——结构化、半结构化或非结构化。现有的基础架构通常会有一些障碍，阻止了这些数据的无缝相关性和全面分析；例如，使用独立的系统来存储和管理这些不同的数据类型。我们也看到了混合系统的出现，但它们往往会让我们失望，因为它们本身并不管理所有数据类型。

在我们看来总觉得奇怪的一件事是，从来没有人肯定很明显的事情: 组织的各种流程不区分数据类型。当您想分析客户支持的有效性时，关于CSR对话的结构化信息(如通话时长、电话结果、客户满意度、调查回复等)和从那次谈话中收集的非结构化信息(如情绪、客户反馈和口头表达的顾虑)同样重要。有效的分析需要考虑到交互的所有组成部分，并在相同的上下文中分析它们，不需要考虑底层的数据是否为结构化数据。一个改变游戏规则的分析平台必须能够管理、存储和检索非结构化数据和结构化数据。它还必须提供探索和分析非结构化数据的工具。

4. 实时分析数据

在活动开展时对其执行分析，为分析型企业提供了一个尚未开发的巨大机会。从历史上看，我们对存储在数据库中的数据运行分析模型和计算。这很适合于在过去几分钟、几小时、甚至几天发生的事件。这些数据库依赖于磁盘驱动器来存储和检索数据。即使是性能最好的磁盘驱动器在实时响应特定事件方面也会有不可接受的延迟。希望提高自己大数据IQ的企业需要能够在生成数据时对数据进行分析，然后采取适当的行动。它的目标是在数据被存储到物理磁盘之前获得洞

察。我们将这种类型的数据称为流式传输的数据，并将所产生的分析称为运动数据分析。根据一天中的时间，

或其他上下文，流式数据的数量可以有很大差异。例如，试想一个数据流承载了交易所中的股票交易。根据交易活动，该流可以迅速膨胀为其正常数量的10至100倍。这意味着，一个大数据平台不仅必须能够支持运动数据分析，并且必须能够有效地扩展，以管理数量不断增长的数据流。

5. 丰富的分析函数库和工具集

大数据平台的主要目标之一是缩短分析周期时间，即发现和转换数据、开发模型并对其评分，以及分析和发布结果所需的时间。我们前面提到过，如果您的平台让您能够速度极快地运行分析，您就拥有了一个支持多次分析迭代和加快模型开发(雪球变大，滚动速度更快)的基础。虽然这是理想的最终目标，但我们仍然需要专注于提高开发人员的生产力。通过轻松地发现数据、开发和部署模型、可视化结果，并集成前端应用程序，您的组织可以让从终端用户(如分析师和数据科学家)在各自的工作岗位上更高效。我们将这个概念称为消费性的艺术。坦率地说，大多数公司都不像LinkedIn或Facebook那样手上有数百名(如果不是数千名)熟练掌握新时代技术的开发人员。消费性是在整个企业中实现让大数据民主化的关键。您不应该只是想，您应该总是要求大数据平台充分利用可加快开发和可视化过程的一组加速器、分析函数库和工具集，让分析时间曲线更平缓。

因为分析是一门新兴学科，数据科学家在创建和可视化模型时有自己的首选机制，这种情况并不少见。他们可能会使用打包的应用程序、使用新兴的开源库，或采用“自己部署”的方法，并使用过程语言构建模型。创建一个限制性的开发环境会降低他们的工作效率。大数据平台需要支持与最常用的分析包的交互，其深度集成有助于将来自那些包的计算密集型活动(如模型评分)推送到平台中。它需要有一组经过开发和测试的大量“并行”算法，可以在大数据上运行。它必须有针对非结构化数据分析的特定功能(如文本分析例程)和用于开发更多算法的框架。它还必须提供以直观、易于使用的方式可视化并发布结果的能力。

6. 集成和治理所有数据源

在过去的几年中，信息管理界在制定完善的数据管理原则方面取得了巨大的进步。这些原则包括数据质量、安全、治理、主数据管理、数据集成和信息生命周期管理的策略、工具和技术。它们在数据中建立真实性和信任，并且对于任何分析程序的成功都非常关键。

大数据平台必须接受这些原则，让它们成为平台的一部分。我们已经多次看到数据质量和治理被认为是“固定”到现有流程的事后想法，这几乎让人感到可怕。我们需要把这些原则作为基础，并且是平台本身固有的。

IBM的大数据和分析战略

在上一节中，我们讨论了新时代的分析如何在组织内推动实现竞争差异化，以及这种差异化如何转化为分析型企业的特定考虑因素。我们也考虑用于提供下一代分析功能的大数据平台应该将什么定义为其主要的迫切需求。在本节中，我们将介绍IBM如何应用这些考虑因素和迫切需求来创建IBM大数据平台。

1. 持续投资于研究和收购

IBM认为，大数据有可能大幅改变组织使用数据和运行分析的方式。业务分析和大数据是IBM的战略赌注，IBM认识到在这一领域证明其领导地位，并为股东创造价值的巨大潜力。

IBM通过持续的投资和战略性收购来推动其对大数据和分析的承诺。2011年，IBM在促进大数据分析的服务和解决方案研发中投资了1亿万美元。看看大数据领域。在过去的五年中，有多少厂商花费超过160亿(截至编写本书时的数据——像大数据的数字一样，当您阅读时，该数据已经过期并且在上升)完成了涵盖30个基于分析的收购？

我们经常发现人们将大数据等同于Hadoop，很坦率地说，这种想法是一个巨大的错误。Hadoop是专为特定任务构建的多种技术之一。大数据的价值(您的C级行政官们真正关心的内容)都围绕着如何将其货币化。可以存储1PB数据很好，但如果您不能分析它，又拥有什么呢？不过是大量的数据。大数据的关键就是分析，并没有其他供应商像IBM一样对分析进行战略投资。IBM最近的两次收购，Netezza(其主要设备已被更名为IBM PureData System for Analytics)和Vivisimo(其产品已被更名为InfoSphere Data Explorer)，已开发了市场领先的创新技术，这些技术现在都已被集成到IBM大数据平台中。IBM还拥有世界上最大的商业研究机构：数百名数学家和科学家在开发领先的分析。最后，IBM拥有世界上最大的专利组合，几乎超过其后四家获得专利最多的公司的总和！

与非结构化数据管理、文本分析、图像特性提取、大规模数据处理有关的许多研究主题和创新都已被纳入IBM的大数据平台。坦率地说，当您考虑有关大数据的战略联盟时，世界上并没有太多其他组织能够像IBM一样提供一个全面的端到端大数据平台。

2. 对开源项目的坚定承诺和生态系统发展的培育

开源社区一直是大数据技术创新的主要驱动力。其中最引人注目的是Hadoop，该软件框架支持数据密集型计算任务的大规模并行处理。Hadoop生态系统包括提供支持实用程序和工具的其他相关开源项目。这些项目提供专业的功能，能够更好地访问在Hadoop分布式文件系统(HDFS)中的数据，方便工作流和作业的协调，支持Hadoop和

其他系统之间的数据移动，实现可扩展的机器学习和数据挖掘算法等。这些技术都属于Apache Software Foundation (ASF)，并且以商业友好的许可模式进行分发。

Apache Hadoop仍然处于其发展的早期阶段。虽然它为大数据提供了一个可扩展且可靠的解决方案，但大多数企业可能会认为它缺少某些特性，缺乏特定的功能，或者需要专门的技能来针对其需求进行调整。

因此，包括IBM在内的技术解决方案供应商都在努力缩小差距，让Apache Hadoop更易于为企业所用。这些技术解决方案供应商可以采取两种不同的方法之一来实现这一目标。第一种方法是以Apache Hadoop的开源代码库为出发点，然后对其进行适当的修改，以解决差距和限制。在软件开发用语中，这一过程被称为分叉。采用这种方法的供应商有效地创建一个特定于供应商的专有Hadoop分发，它有点封闭，并且与社区对开源组件所应用的创新和改进相隔离。但却造成与其他互补技术的互操作性困难得多。

第二种方法是原样保留开源Apache Hadoop组件，不修改代码库，而是添加其他层和可选组件来增强并丰富开源分发。IBM在其InfoSphere BigInsights (BigInsights)产品中采取了第二种方法，该产品将Apache Hadoop的开源组件作为一个“核心”层，并围绕它构建增值组件。这让IBM能够在其开发中快速采取对核心开源项目的任何创新或变更。这也让IBM很容易认证那些与开源Apache Hadoop集成的第三方技术。

这种将其他基于开源的Hadoop发布纳入自己产品的模块化战略，让IBM能够保持开源组件的完整性，并解决它们的局限性。BigInsights是IBM认证的Apache Hadoop版本。此外，BigInsights所提供的许多增值组件(如BigSheets和Advanced Text Analytics Toolkit)支持在Hadoop组件的Cloudera分发上使用。(其他分发目前正在考虑中。)有时候，我们对IBM在开源领域的承诺提出质疑。这种让我们猝不及防的问题，并不是因为问题很难——相反，我们感到措手不及是因为我们意识到，也许IBM对它在这个领域成就没有进行足够好的传播工作，因此，我们决定在这里要做这件事(尽管是简短的)。郑重声明，IBM100%致力于开源，并且IBM 100%致力于Hadoop。

我们的许多工程师向Apache Hadoop开源项目及其相关的生态系统提供代码片段(在开源世界中，他们被称为提交者)。IBM在发明技术并为开源社区进行捐赠方面有着悠久的历史，例子包括Apache Derby、Apache Geronimo、Apache Jakarta、DRDA、XERCES；这样的例子不胜枚举。事实的开源工具集Eclipse来自IBM。文本分析是围绕大数据的一种主要用例，而IBM贡献了非结构化信息管理架构(UIMA)。搜索是一个大数据的先决条件推动者，IBM是Lucene的一个

主要的贡献者(在电视节目Jeopardy!中获胜的Watson技术所演示的搜索技术)。在这本书中，我们没有足够的篇幅详述IBM对开源的承诺，但我们认为已经表达了我们的观点。

IBM在大数据领域还构建了强大的解决方案供应商生态系统。目前，其合作伙伴(包括在IBM大数据平台上接受培训以及获得认证的技术供应商和系统集成商)的数量达到了三位数。

3.支持大数据的多个切入点

大数据技术可以为组织解决多个业务问题。因此，组织往往可以抓住最好的采用方法。作为从业者，我们有时会看到IT组织在启动大数据措施时就好像是在进行一个寻找问题的科学实验；我们将告诉您，以我们的经验来看，缺乏重点和不明确的预期通常会导致不利的结果(这对于工作安全性没有好处)。在我们看来，大部分成功的大数据项目都从清晰识别业务问题或难题开始，其次是应用适当的技术来解决这个问题。毕竟，我们还没有看到一个客户因为喜欢软件的外观就对它感到兴奋。(虽然我们认为Netezza绝对是让硬件看起来很酷的第一个产品。)在正确的地方开始至关重要。事实上，第一个大数据项目的成功可以决定这一技术在组织中其他地方的采用有多广泛和多迅速。

我们已经识别了在我们的客户协议中所遇到的一些最常见难题(它们可有效地作为大数据项目的“触发器”)。出于这个原因，我们认为IBM大数据平台有五个大数据切入点，如图3-2所示。

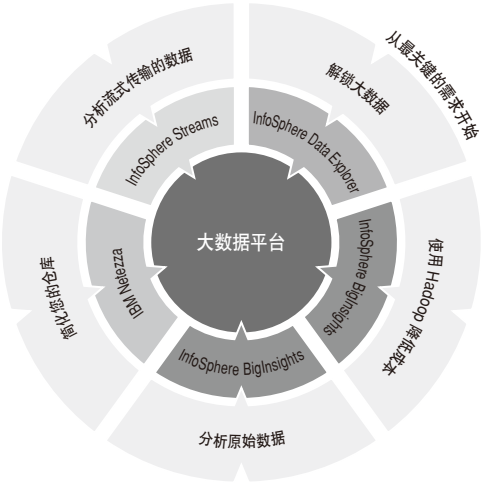


图3-2 IBM的大数据平台支持多个切入点。

随着您从一个切入点移动到第二个和第三个项目，IBM大数据平台将提供可量化的效益，因为它构建于一组在核心集成的共享组件之上。我们发现的一个典型结果是，在一个大数据项目中使用的资产不仅增加下游项目成功的机会，还可以加速其交付。

例如，您可能为应用程序开发一个情感分析包，为服务台应用程序处理数百万个电子邮件。随后的项目可以获得此资产(在静止时收获的数据)，并透明地将其部署到一个运动数据应用程序，评估实时Twitter消息流的情感趋势。

最后，值得一提的是，IBM大数据平台并不是一个全有或全无的命题。恰恰相反: 您可以从单个产品、产品的一个子集或整个平台开始，为您提供了成功交付首个项目所需的灵活性和敏捷性，然后以那里为起点逐渐“提升”您的大数据IQ。在本节的余下部分，我们会讨论每一个IBM大数据平台的切入点。

解锁大数据

您有多少次无法在自己的笔记本电脑上找到明明知道肯定存在的内容？您有多少次在寻找一个文件时却偶然发现一个已经完全忘记的文件？如果您无法真正处理位于自己笔记本电脑上的数据资产，想象大型企业的这个难题会造成什么后果！说实话，我们发现企业常犯的错误是不知道自己可能已经知道了什么，因为他们不知道他们有什么。换句话说，他们有数据，但不能了解它。这是与“解锁大数据”有关的难题。

对于要访问多个数据源，但没有基础架构将所有数据放在一个中央位置，或者没有资源可以开发分析模型，从数据中获取洞察的组织来说，这个问题很复杂。在这些情况下，最关键的需求可能是迅速解锁驻留在这些数据中的价值，而无需将数据移到任何地方，并且在以信息为中心的新应用程序中使用大数据源。此类实现可以产生显著的商业价值，从减少用手动工作去搜索和检索大数据，到更好地了解现有的大数据源，然后再到分析。投资回收期通常都很短。此切入点让您能够以联合的方式发现、导航、浏览、查看和搜索大数据。

使用Hadoop降低成本

组织在降低其数据仓库的整体成本方面可能会遇到一个特定的难题。他们可能会保留极少使用但占据宝贵存储容量的某些数据组。这些数据组可能是成本较低平台的扩展候选原因(有时也被称为可查询档案)，尽管该平台的性能较差，但它提供了存储和检索功能。此外，某些操作(如转换)可以被卸载到更为经济高效的平台上，以提高ETL(提取、

转换和加载)流程或数据仓库环境的效率。此类型问题的切入点可能从静止数据引擎Hadoop开始。此处的主要价值创造领域是节约成本。通过有选择地将工作负载和数据集推送到Hadoop平台上，组织能够保留他们的查询功能，并针对合适的数据和工作负载充分利用Hadoop经济高效的处理功能。

分析原始数据

有些情况可能是企业想扩大其数据仓库的价值，方法是纳入新的数据类型，并推动实现新的分析类型。他们的主要需求可能是分析来自一个或多个源的非结构化数据。他们可能希望对原始格式的数据进行分析，以克服将非结构化数据源转换为结构化格式的过高成本。一个基于Hadoop的分析系统对于此类问题可能是合适的切入点。企业往往可以使用这种方法获得显著的价值，因为他们可以解锁以前未知的洞察。这些洞察非常关键，有助于保留有价值的客户、识别以前未被发现的欺诈行为，或发现操作流程中可以改变游戏规则的效率。

简化您的仓库

我们经常遇到的情况是，业务用户受阻于通用的企业级数据仓库中性能较差的分析，因为他们的查询需要花几个小时来运行。与提高数据仓库的性能相关的成本高得让人望而却步。企业可能需要简化其仓库，并让它快速地启动和运行。在这种情况下，转移到专用的大规模并行系统和分析设备可能是大数据的完美切入点。许多组织通过使用我们的技术实现了10 - 100倍的深度分析性能提升，并降低了拥有成本，提高了员工的工作效率。

分析流式传输的数据

组织可能有多个流式传输的数据源，并希望快速处理和分析易变的数据，实时采取行动。他们往往根本无法充分利用这些数据，因为在分析之前可能需要收集和存储太多的数据。或者，与将数据存储在磁盘上，然后对其进行分析的延迟对于他们要推动的实时决策来说可能是无法接受的。利用流式传输数据，并将其转化为可操作洞察的能力可能是另一个大数据的切入点。其好处将是能够制定实时决策，并通过分析运动中的数据和只存储必要的数据来实现成本节省。

一个灵活的、基于平台的大数据方法

在本章的最后，我们将可用的服务从IBM大数据平台(如图3-3所示)映射到提供服务的产品，并提供指向在本书其他部分中更多详细信息的指针。

图3-3 IBM大数据平台

1.可视化与发现:

由速度驱动的IBM InfoSphere Data Explorer对于需要了解其数据源的范围和内容的组织，IBM InfoSphere Data Explorer (Data Explorer)是一个很好的起点。该产品让企业能够通过联合的搜索、发现和导航，解锁在企业内外所拥有的数据。它让每个人(从管理层到知识型员工，到一线员工)都能够在一个视图中访问他们需要的所有信息，不需要考虑数据的格式或位置。他们不需要浪费时间分别访问每个筒仓，Data Explorer让他们能够跨所有可用的来源发现数据并无缝地进行导航，并提供了跨存储库的可见性这一额外优势。它保护信息，让用户只看到允许他们看到的内容，就像他们直接登录到目标应用程序一样。此外，Data Explorer让用户能够对内容进行评论、标签和评分，以及为他们希望与其他用户共享的内容创建文件夹。然后，所有这些用户反馈和社交内容都被反馈回Data Explorer的相关性分析功能，确保向用户显示最有价值的内容。宝洁等企业使用这种技术已经能够为员工提供对30多个数据存储库的可见性，从而简化了他们的支持和运营工作。

Data Explorer是开始大数据之旅的一个不错起点，因为您可快速发现和检查自己手上的数据资产，以确定下一步需要平台的哪些其他部分。我们将在第7章讨论这项技术。

2. Hadoop系统: IBM InfoSphere BigInsights

Hadoop对于希望在一个地方结合多种数据类型(包括结构化和非结构化)进行深度分析的组织来说是一项理想的技术。它还让组织能够通过卸载数据和工作负载，从而降低其数据管理基础架构的成本。IBM的InfoSphere BigInsights构建于开源Hadoop之上，并使用企业所需的必备功能增强它。其优化可以自动调优Hadoop工作负载和资源，实现更高的性能。它有一个直观的电子表格风格的用户界面，让数据科学家们可以快速检查、浏览和发现数据关系。它提供了安全性和治理功能，确保敏感数据得到保护和保障。BigInsights预包装了开发工具，让技术团队能够更轻松地创建应用程序，而不必首先经过详尽的培训来成为Hadoop专家。在第5章中，我们将介绍IBM的Hadoop分发。

3.流计算: IBM InfoSphere Streams

特别无法存储分析所需的足够的数据时，流计算让组织能够立即应对不断变化的事件。它还让组织能够更高效地预筛选并有选择性地存储高速数据。IBM InfoSphere Streams (Streams)让组织能够实时地分析流式传输的数据，从而提供了这样的能力。Streams采用模块化设计，具有无限的可扩展性，每秒可以处理数百万个事件。它能够同时分析多种数据类型，并实时执行复杂的计算。我们将在第6章详细描述Streams和运动数据的分析。

4. 数据仓库设备: 由Netezza技术驱动的IBM PureData System for Analytics

我们经常会发现企业在艰难地对付其数据仓库环境的复杂性。他们的数据仓库往往充满了数据，并且不适合于一个特定的任务。获得深入的分析洞察可能过于复杂或非常昂贵。IBM通过基于分析的IBM PureData System解决这个难题。IBM PureData System for Analytics(最新一代Netezza设备的新名称)是一个专用设备，适用于大量结构化数据上的复杂分析工作负载。它的设计考虑到了简单性，它只需要最少的管理，不需要性能调优。它使用独特的硬件辅助的查询处理机制，让用户能够以极快的速度运行复杂的分析。我们将在第4章讨论这项技术。

5.分析加速器

IBM基于平台的方法的目标之一是加快大数据项目实现价值的速度。IBM大数据平台将预构建的分析、可视化和行业特定的应用程序打包为平台的一部分，从而实现这一目标。分析加速器包含一个预构建的函数库，使用适当的引擎在数据驻留的地方以其原生格式分析数据: InfoSphere BigInsights、InfoSphere Streams或一个基于分析的IBM PureData System。预构建的函数库包括

处理文本、图像、视频、声学、时间序列、地理空间和社交数据的算法。函数范围涵盖数学、统计、预测和机器学习算法。加速器还包括相应的开发工具，让用户能够为大数据平台定制开发分析应用程序。我们会在第8章介绍文本分析，并在第9章讨论三个大数据分析加速器，其作用是让您快速推进大数据项目。第5章会介绍发现工具(如BigSheets)以及BigInsights产品。

6. 信息集成和治理

IBM大数据平台包括为多个数据源专门构建的连接器。它有针对常见的文件类型、数据库，以及非结构化数据和流式传输数据源的适配器。它的各个数据处理引擎之间也已实现深度集成，可以在基于分析的IBM PureData System、InfoSphere BigInsights和InfoSphere Streams之间无缝移动数据。

安全性和治理是大数据管理的关键方面。大数据平台可能包含需要保护的敏感数据、需要执行的保留策略，以及需要治理的数据质量。IBM公司拥有信息生命周期管理、主数据管理、数据质量和治理服务的强大产品组合。这些服务是集成的，可在大数据平台中使用。我们会在第10章和第11章讨论这个主题。

小结

我们在本章开始时描述了任何企业的基础组件。我们相信，这些核心功能是任何大数据平台的DNA——无论您是在考虑该领域中的IBM产品还是其他供应商的产品，我们丰富的经验表明，这就是提高您的大数据IQ所需要的。我们的大数据宣言构建于这些基础组件之上。

我们在本章的最后向您介绍了IBM大数据平台，该解决方案为您提供了一个灵活的大数据采用计划，让您可以应对特定的业务挑战。该平台的真正好处是您将来自一个实施的新功能应用到下一个实施时能够充分利用可重用的组件(分析、加速器和策略)。该平台让您能够使用单一集成平台来管理所有企业数据，提供多个数据引擎(每个都针对具体的工作负载进行了优化)，并使用一致的工具和实用程序来操作大数据。在平台内的预集成组件也减少了实施时间并降低了成本。IBM是提供这一广泛和均衡的大数据视图以及大数据平台需求的唯一供应商。

IBM通过“乐高积木块”的方式交付其大数据平台。例如，如果客户在现有的数据仓库方面遇到了挑战，并且无法处理大规模结构化数据，其最初的起点将是一个专用的数据仓库设备。以后可以将其扩展为包括用于分析原始静止数据的Hadoop，也可以包括用于对运动中的数据进行分析的流计算。最后，IBM提供了各种独特的功能来应对各种难题。我们相信，目前市场上还没有具备同等完整性、可消费性和功能性的大数据平台。

分析静止的大数据

4 一个提供高性能深入分析的大数据平台: IBM PureDataSystems

数据仓库系统和技术旨在为组织提供按需访问信息的能力，帮助他们更快地对信息做出反应，推动更快地制定决策。诚然，用于描述数据仓库大小的语言已从GB变为TB和PB，工作负载已从主要包括操作性工作负载变得越来越侧重于分析，描述并发性的数字也从几百变为几千。这些变化让许多人认识到，无法扩展传统的数据仓库方法来满足如今的挑战。

组织期望在他们迫切需要关键信息时，提供该信息的平台应是他们最不需要担心的。毕竟，打开房间的灯时，您是否会考虑让灯发光的布线和电路？显然不会。您期望能够轻松地使用所需的工具来完成工作(当然，您必须拨动电灯开关，但仅此而已)。为了支持如今数据仓库越来越高的复杂性、并发性和数量，IT组织需要一个简单、可靠且能够轻松地处理几乎任何工作负载的平台。

第一代数据仓库技术是仿照运行于大型对称多处理(SMP)机器之上的基于OLTP的数据库来实现的。这些机器的一些内在架构限制使其无法成为可行的分析平台。后面几代技术尝试将并行处理技术和分布式存储子系统合并到该架构中。这带来的操作复杂性(各种类型的索引、索引上的索引、优化提示等)让这些系统的操作变得更加复杂，维护费用更加高昂。

在越来越高的数据量和多样化的工作负载上实现一致的性能，而不显著增加总体拥有成本(TCO)，这始终是数据仓库技术中的最大挑战。所有数据仓库操作的最大瓶颈始终是数据库引擎从磁盘读取数据和向其写入数据的速度，这也称为磁盘I/O瓶颈。

当仓库由多个部分组成时，各种提供商提供了许多I/O创新来尝试解决这一瓶颈；但是，这些创新被独立地带入市场，在各个仓库层级表现出极少的协同性：关系型数据库管理系统(RDBMS)、存储子系统和服务器技术。例如，在存储子系统中使用缓存，服务器上更快的网络结构，以及软件优化(如分区和索引)都是为了最大限度减少磁盘I/O而推向市场的优化技术。而且，尽管它们解决了局部的问题并且提供了些许帮助，但这些优化集中起来未能显著改善磁盘I/O。此外，这些优化技术依靠数据仓库设计人员来提前预测查询模式和检索需求，以便它们可调优系统性

能。这不仅在面对新的报告和分析需求时影响了业务敏捷性，还需要大量的人力来维护、调优和配置各个数据仓库层。结果是，这些系统全都变得难以维护且管理费用高昂。

IBM PureData System for Analytics设备(以前称为IBM Netezza Data Warehouse Appliance，常被称为Netezza)专为克服这些具体的挑战而开发。(因为本章介绍IBM PureData System设备的起源，也就是Netezza技术的历史，所以本章后面会将该设备和技术统称为Netezza)。事实上，可以这么说，Netezza引发了数据仓库中的设备革命，将数据库、处理和存储组件集成到一个灵活、紧凑、特制且已优化的分析工作负载系统中。这个创新平台提供了行业领先的性价比和设备简单性。作为一种为高速大数据分析而特制的设备，它的强大并非源于最强大和昂贵的组件(这种会让它的成本效益曲线变得更陡)，而源于恰当组件的组装方式和它们为实现最高性能而进行的协同工作。简言之，我们的目的不是构建一台超级计算机，而是构建一个设计精美的系统，通过无与伦比的分析执行能力来解决常见的瓶颈。为此，Netezza将大规模并行处理(MPP)流和多核CPU与Netezza独有的硬件加速功能(我们称之为Netezza的秘密武器)相结合，目的是最大限度减少磁盘I/O瓶颈，提供更为昂贵的系统永远无法匹敌，甚至无法接近的性能和专家调优功能。我们想阐述一下上句中的“流”这个词。在此上下文中，我们谈论的是将数据从磁盘上的数据库组件流式传输到内存中，以供处理。第6章中将介绍表示使用InfoSphere Streams对运动中的数据执行持续处理的流。

一定要注意，Netezza没有借用具有已知不足的旧系统并使它与一个存储层保持协调。它是从头构建的，专门用于在大量结构化数据上运行复杂的分析。作为一个易于使用的设备，该系统开箱即用地带来了它的非凡成果，无需进行索引或调优。设备简单性也延伸到了应用程序开发上，支持实现快速创新以及将高性能分析带给最广泛的用户和流程的能力。对于用户及其组织，这意味着为所有需要的人提供了最佳的智能，即使所有方面的需求都在增加也是如此。

回顾Netezza的历史和如今的大数据时代，我们认为可以合理地断言，Netezza将CFO/CIO讨论从“花钱来省钱”转变为“花钱来赚钱”：它为各种规模的企业提供了全面实现深入分析的能力，让曾经与这一领域相关的成本曲线变缓了。

本章将比前面几章更多地介绍技术。这可让我们能够更好地解释概念，演示使这项行业领先的Netezza技术如此特别的“秘密武器”。

Netezza的设计原则

Netezza的数据分析方法已获得专利且久经考验。它的目标始终是最大限度减少数据移动，用“物理速度”、并行地且大规模地处理它——所有这些都在一个低成本、容易使用的设备中提供。Netezza设备自诞生以来已经历了多次更新(它的最后一次更新包括将名称改为IBM PureData System for Analytics)，但它的架构始终基于一组核心的设计原则。这些设计原则已成为Netezza行业领先的性价比标志，本节将介绍这些原则。

设备简单性: 最大限度减少人力劳动

企业花费越来越多的资金来请人管理他们的系统。现在，在这样的环境中想想这一结果，其中可通过一个全球分布式交付模型提供更廉价的劳动力，您可看到所需的人力劳动量是一个问题。

设备是针对一种具体用途而优化的专用装置。它们是自成一体的，随时可用。您可非常快地安装它们且易于操作。它们拥有标准的接口，因此其他系统能够与它们互操作。所有这些特征让设备成为一个提供IT解决方案的非常优秀的工具。在一些领域，如网络和存储解决方案，设备是提供具体功能的首选方式。

Netezza在数据仓库和分析领域开创了设备的概念。它的所有技术都以设备的形式实现，为最终用户屏蔽了平台的底层复杂性。只要存在设计，就需要权衡简单性与设备的任何其他方面。不同于其他解决方案，设备始终能运行，能够以极高的速度处理要求很高的查询和混合工作负载。甚至通常比较耗时的任务都得到了极大简化，如安装、升级及确保高可用性和业务连续性，这节省了宝贵的时间和资源，减轻了操作风险。

硬件加速: 在靠近数据存储的地方处理分析

Netezza的架构基于计算机科学的一条基本原则: 操作大型数据集时，除非绝对需要，否则不要移动数据。将大型数据集从物理存储单元移动到计算节点会增加延迟并影响性能。Netezza使用创新性的硬件加速功能最大限度减少了数据移动；例如，它使用现场可编程逻辑门阵列(FPGA)在数据流中尽早地滤出不相关的数据，基本上与数据传出磁盘的速度一致。这个在数据源附近消除数据的流程去除了各种I/O瓶颈，让下游组件(如CPU、内存和网络)无需处理过多的数据；这会在系统性能上带来显著的乘数效应。

平衡、大规模并行架构: 提供线性可伸缩性

Netezza架构的每个组成部分(包括处理器、FPGA、内存和网络)都经过精挑的细选和优化，能以磁盘的物理结构所允许的最快速度提供数据，同时最大限度降低成本和功耗。

Netezza软件将这些组件设计为以一种管道方式同时操作数据流，进而最大限度提高了利用率并从每个MPP节点获取了最高的吞吐量。使用开放、基于刀片的组件让Netezza能够非常快地整合各种技术增强，而FPGA的涡轮增压效应(一种平衡的硬件配置)和紧密耦合的智能软件相结合，提供了比各个元素高得多的总体性能。事实上，Netezza自诞生以来，其性能每两年都会增长4倍，远远高于其他信誉卓著的供应商产品。

模块化设计: 支持灵活的配置和极高的可伸缩性

在传统的基于设备的架构中，一个重要问题是它们在数据量的增长超过设备的物理容量之后的扩展能力。Netezza通过一种模块化设备设计解决了这个问题，该设计能轻松地从数百GB扩展到数十PB的可查询用户数据。事实上，该系统的设计具有很高的适应能力，能满足数据仓库和分析市场中不同领域的需求。使用开放的、基于刀片的组件让您可在配置中轻松修改磁盘-处理器-内存比率，以迎合以性能或存储为中心的需求。同一种架构还支持基于内存的系统，这些系统为任务关键型应用程序提供了极快的实时分析。本章剩余部分将深入分析这些设计原则如何转换为实践。

内部结构如何? Netezza设备架构概述

Netezza架构基于一种非对称大规模并行处理(AMPP)方法，将对称多处理(SMP)的最佳要素与MPP相结合，打造了一款特制的设备，可在PB级数据上运行极快的分析。

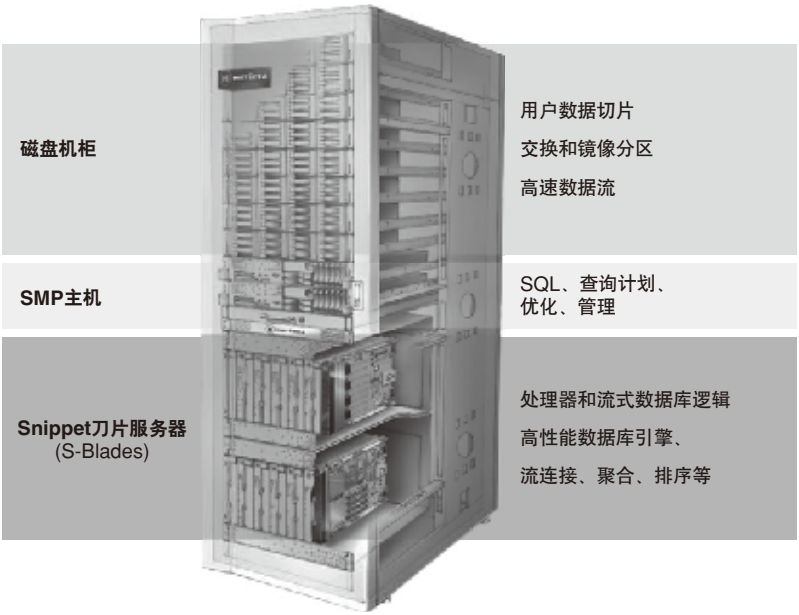


图4-1 IBM Netezza设备(现在称为IBM PureData System for Analytics)

一个单机架Netezza系统如图4-1所示。可以看到，它有两个高端机架挂载式服务器，称为主机，它们同时用作Netezza设备的外部接口和大规模并行基础架构的控制器。主机通过一个内部网络结构连接到一组snippet刀片服务器(我们常常将它简称为S-blade)，其中执行主要的数据处理工作。S-blade通过一个高速互联设备连接到一组存储数据的磁盘机柜。

如果任何组件发生故障，主机、磁盘或S-blade，会自动执行恢复工作。丰富的自动故障检测和恢复选项证明了设备方法的优势。Netezza拥有内置的组件和操作冗余性，该系统会自动且无缝地响应关键组件的故障。

Netezza设备内部结构分析

Netezza的AMPP架构的每个组件(包括处理

器、FPGA、内存和网络)都经过了精挑细选和集成，以实现一个平衡的总体系统。除了出色的性能，这种平衡的架构还可提供线性可伸缩性，可扩展到超过1000个并行执行的处理流，同时具有非常经济的TCO。这些并行流可以同“分步解决”工作负载协同工作。让我们看看一个Netezza系统的关键组件，如图4-2所示。

Netezza主机

主机是系统的主要接口。主机是高性能的Linux服务器，设置为一种主动-被动配置以实现高可用性(注意，图4-2中有两个主机堆叠在顶部)。主机将SQL查询编译到称为代码段的可执行代码片段中，创建优化的查询计划，以及将代码段分发给MPP节点来执行。活动的主机为外部工具和应用程序提供了一个标准化的接口，如商业智能报告工具、数据集成和ETL工具、高级分析包、备份和恢复工具等。这些外部应用程序使用普遍的协议(如JDBC、ODBC和OLE DB)来连接Netezza主机。

主机同时运行着高可用性(HA)集群软件和镜像技术。所有向活动主机的写入操作都镜像到备用主机，因此它可在任何时候发生故障时接管系统的处理工作。备用主机通过以太网链接监视活动主机，在活动主机发生故障时承担主要角色。

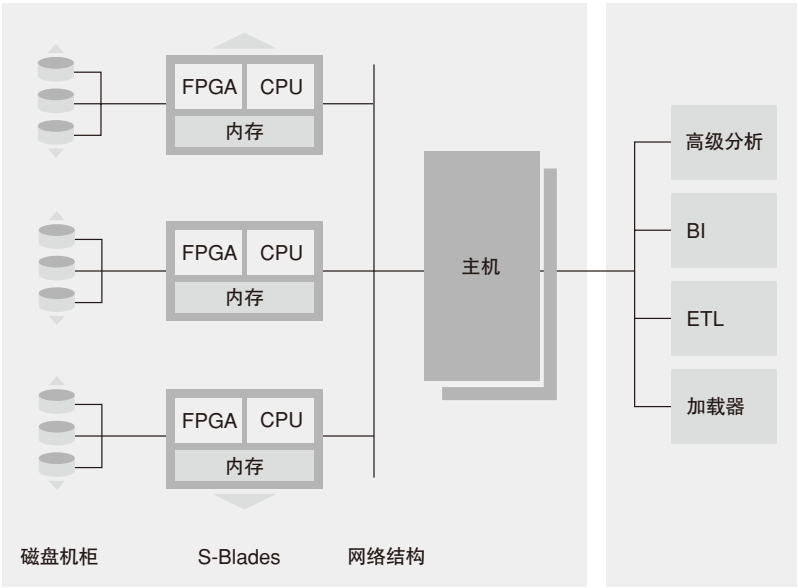


图4-2 Netezza AMPP架构

S-blade

S-blade是组成Netezza设备的MPP引擎的智能处理节点。每个S-blade是一个独立的服务器，包含强大的多核CPU、多引擎FPGA和GB级RAM，所有这些组件都均衡地且并发地工作，从而提供最高的性能。

Netezza也可防御S-blade故障。系统管理软件持续监视每个S-blade，如果它检测到故障，S-blade会中断服务并将其处理负载自动传输到备用的S-blade。S-blade的监视包括刀片内存上的错误更正代码(ECC)检查。如果系统管理软件检测到ECC错误超出了故障阈值，它会让S-blade中断服务。

磁盘机柜

系统的磁盘机柜(每个机柜包含一个表的数据切片)包含受RAID保护的高密度、高性能磁盘。磁盘机柜通过高速互联接口连接到S-blade，支持Netezza中的所有磁盘以最大速率同时将数据传输到S-blade。

从机箱中的每个S-blade到每个磁盘有两条独立的数据路径。使用RAID 1配置建立每个驱动器的镜像。发生磁盘故障时，处理工作会在镜像磁盘上完成而不会中断任何服务。I/O会由存储子系统重定向到镜像驱动器。每个设备中包含备用驱动器，所以它可选择一个备用驱动器来代替故障驱动器并重新生成其内容来恢复冗余，从而透明地实现自我修复。

网络结构

整个Netezza MPP网格中的节点间通信在一种网络结构上进行，该网络结构运行一种基于IP的定制协议，该协议充分利用了这种结构的总体典型带宽并消除了拥塞，即使具有持续的爆炸式网络流量也是如此。该网络经过了优化，能扩展到超过1000个节点，并且允许每个节点同时向其他每个节点执行大规模的数据传输。

Netezza的定制网络协议专为与高容量数据仓库相关的数据量和流量模式而设计。它可确保最大限度地利用网络带宽，同时不会让它负担过重，因此可实现接近网络的数据传输速度的可预测性能。

Netezza系统中的流量在3个不同区域内流畅地流动:

- 在广播模式中，从主机到代码段处理器(1比1000以上)
- 从代码段处理器到主机(1000比1)，在S-blade中和系统机架级别上执行聚合
- 在代码段之间(1000比1000)，数据大规模地自由流动，以进行中间处理

Netezza网络结构基础架构拥有与其他主要组件相同的冗余级别。事实上，每个Netezza设备拥有两个完全独立的网络，而且因为每个冗余主机连接到一个冗余网络，所以网络结构任何一部分的故障都可通过切换到备用主机来克服。数据网络不仅是冗余

的，而且在物理上与各个管理网络分开，这让系统能够评估其组件的健康状况，即使数据网络遇到了问题也是如此。

秘密武器: FPGA协助的分析

FPGA是Netezza设备的性价比优势的一个关键推动因素。FPGA是一个半导体芯片，它配备了大量内部逻辑门，可编程这些逻辑门来实现几乎所有逻辑功能。对一个FPGA进行编程后，它的操作更像是特制的硬件，而不是一般用途的处理器，因为它能最佳地执行很小范围的任务。FPGA在管理特殊用途的流处理任务上特别有效，被广泛用于数字信号处理、医疗成像和语音识别等应用程序。实际上，可能的情况是您已有一个FPGA但您却不知道！您看过DVD或蓝光碟片吗？它们背后可能有一个FPGA在帮助从旋转的碟片上平稳地读取高质量压缩数字数据。FPGA的尺寸很小(大约1″ ×1″ 的方形硅片)，但能极为高效地执行已编程的任务，同时只需极少的功耗，发热量也极少。

图4-3演示了每个S-blade中启用了FPGA的处理功能。来自存储阵列的一个专用的高速互联接口让数据能以从磁盘传出的速度传送到内存。压缩的数据使用一种智慧算法缓存在内存中，这可确保可从内存中提取最常访问的数据，无需再访问磁盘。

每个FPGA包含在数据流上执行过滤和转换功能的嵌入式引擎。可动态配置这些引擎(可通过软件修改或扩展它)，可通过在执行查询期间提供指令为每个代码段进行定制，而且以极高的速度处理数据流。您在图4-3中的FPGA方框中看到的所有引擎都会并行地运行，能够以物理速度解压缩并滤出95%到98%的表数据，仅保留与查询相关的数据。

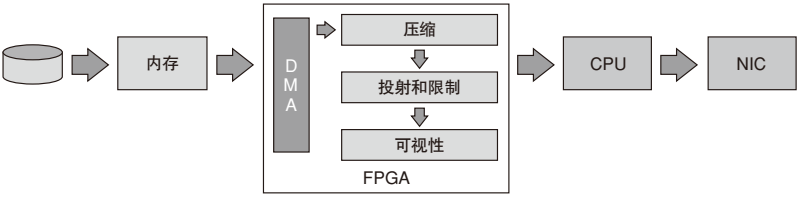


图4-3 Netezza中FPGA协助的代码段处理

图中所示的CPU核心也并行运行，并发地处理流中的剩余数据。我们简化了整个图，使其更容易理解，但这里介绍的流程会在设备中运行的约100个并行代码段处理器上重复运行。在一个包含10个机架的系统上，这可能代表着1000个并行代码段处理器，性能比贵得多的竞争系统要高出几个数量级。FPGA包含以下嵌入式引擎:

- **压缩引擎** Netezza中的所有数据都以压缩形式存储在磁盘上。压缩引擎位于FPGA内，它能以网络的传输速度解压数据，迅速将磁盘上的每个数据块转换为内存中的4到8个数据块。使用基于半导体的技术而不是软件来实现数据的压缩和解压，这将系统性能提高了4到8倍，显著加快了任何数据仓库中最缓慢的组件(磁盘)的处理速度。
- **投射和限制引擎** 此引擎通过基于SQL查询的SELECT和WHERE子句中的参数来滤出列和行，进一步提高了性能。
- **•可视性引擎** Netezza在流速度上保持着ACID(原子性、一致性、隔离性和耐久性)合规性，使用它的可视性引擎来滤出不应被查询“看到”的行(如属于一个尚未提交的事务的行)。

Netezza中的查询协调

所有查询和分析都通过主机机器进入Netezza，优化器、编译器和调度程序在主机中将它们分解为许多不同的代码段，这些代码段的指令又被进一步分发到多个S-blade中(然后所有S-blade同时对它们在本地管理的数据切片处理工作负载)。本节将使用图4-4更详细地讨论各个步骤。

查询优化

Netezza优化器的智能性是其最大的优势之一。其设计是为了创建专门针对Netezza的AMPP架构而优化的查询执行计划。

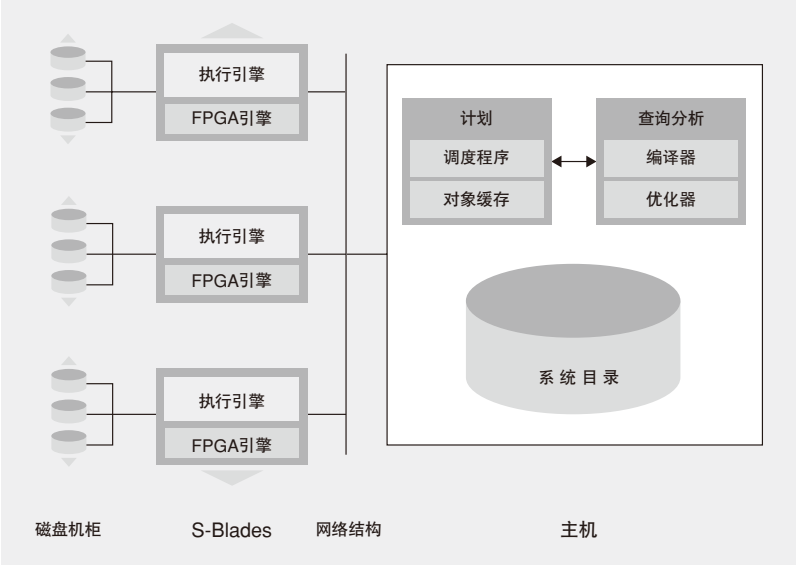


图4-4 Netezza中的查询协调

优化器使用系统中的所有MPP节点收集一个查询中引用的每个数据库表的详细最新统计数据。其中大部分度量指标都在查询执行期间以极低的开销捕获，得到的是特定于每个查询的即时统计数据。在处理开始之前，优化器使用这些统计数据转换查询，这有助于最大限度减少磁盘I/O和数据移动。优化器执行的典型转换操作包括：确定正确的连接顺序，重写表达式和通过SQL操作删除冗余。

Netezza的设备本质与能够彼此通信的集成组件相结合，使其基于成本的优化器能够更准确地度量与一个操作相关的磁盘、处理和网络成本。通过依靠准确的数据，而不是单纯的启发法，优化器能够生成以极高效率利用所有组件的查询计划。优化器智能的一个示例是，它能确定一个复杂连接中的最佳连接顺序。例如，将多个小表连接到一个大的事实表时，优化器能够选择将小表完整地广播到每个S-blade，同时保持让这个大表分散在所有代码段处理器上。此方法最大限度减少了数据移动，同时充分利用AMPP架构实现连接的并行化。

代码段编译

编译器将查询计划转换为多个可执行的代码片段，称为代码段。查询片段由代码段处理器跨设备中的所有数据流并行地执行。每个代码段有两个元素：由各个CPU核心执行的已编译代码，以及一组定制该代码段的嵌入式引擎过滤的FPGA参数。这种逐个代码段的定制方式让Netezza能够有效提供一种能针对各个查询动态地优化的硬件配置。

主机使用一种称为对象缓存的特性来进一步加速查询性能。这是之前已编译的代码段的大型缓存，支持参数变化。例如，一个具有子句WHERE NAME=′ Bob′ 的代码段使用与具有子句WHERE NAME=′ Jim′ 的代码段相同的已编译代码，但使用可反映不同名称的设置。此方法消除了99%以上的代码段的编译步骤。

即时调度

Netezza调度程序在复杂的工作负载间均衡地执行操作，以满足不同用户的目标，同时保持最高的利用率和吞吐量。它考虑了许多因素，包括查询优先级、大小和资源可用性，以确定何时在S-blade上执行代码段。设备架构让调度程序能够从系统的每个组件收集有关资源可用性的最新、更准确的度量指标。复杂的算法是这个调度程序的核心，让Netezza能够最大限度提高系统吞吐量，利用近乎100%的磁盘带宽，同时确保内存和网络资源的负担不会过重以及系统效率降低。Netezza设备的这一重要特征可确保系统保持以峰值吞吐量运行，甚至在非常高的负载下也是如此。调度程序提供“绿灯”时，会将代码段广播到整个智能网络结构的所有代码段处理器。

大规模并行代码段处理

每个S-blade的代码段处理器收到特定的指令，告诉它需要执行代码段中的哪一部分。除了主机调度程序，代码段处理器还拥有自己的智能抢先调度程序，允许来自多个查询的代码段同时执行。这个调度程序会考虑查询的优先级和为发出查询的用户或组所保留的资源，决定何时调度一个特定的代码段来执行和执行多长时间。以下步骤列出了与代码段处理相关的事件序列:

1. 每个代码段处理器上的处理器核心使用查询代码段中包含的参数来配置FPGA引擎并设置一个数据流。

2. 代码段处理器从磁盘阵列将表数据读入内存。它还会在访问磁盘中的数据块之前质询缓存，避免在数据已位于内存中时执行扫描。代码段处理器使用一种名为ZoneMap加速的Netezza创新来减少磁盘扫描。

   ZoneMap加速利用了数据仓库中各个行的自然顺序，可将性能加快几个数量级。该技术避免了扫描其列值在查询的开始和结束范围外部的行。例如，如果一个表包含两年的每周记录(约100周)，并且一个查询仅查找一周的数据，ZoneMap加速可将性能提升100倍。与传统数据仓库技术中的优化相关的典型索引不同，会自动针对每个数据库表创建和更新ZoneMap，同时不会产生任何管理开销。

3. FPGA在数据流上执行操作。首先，它通过以网络的传输速度解压数据流，将速度提高4到8倍。接下来，它的嵌入式引擎滤出任何与查询无关的数据。剩余的数据被传送回内存供CPU核心并发地处理(如图4-3所示)。这一阶段得到的数据通常只是原始流的极小部分(2%到5%)，大大减少了处理器核心所需的执行时间。

4. 处理器核心挑选数据流并执行核心数据库操作，如排序、连接、聚合等。它还应用嵌入到代码段处理器中的复杂算法来执行高级分析。然后在内存中组装每个代码段处理器工作的结果，以生成整个代码段的子结果。超过1000个代码段处理器可同时工作，可并行执行数百或数千个查询代码段。代码段处理器使用智能网络结构来与主机(和彼此)通信，并执行中间计算和聚合。一些高度复杂的算法(如矩阵计算)需要在节点之间通信。出于此原因，Netezza设计了一个消息传递接口来传达中间结果和生成最终结果。

5. 最后，主机对从代码段处理器收到的中间结果进行组装，编译最终的结果集，然后将它返回给应用程序。当然，在此期间，其他查询会在完成各个阶段时流经系统。这为Netezza提供了另一个优化点: 因为原始的已压缩数据块仍在内存中，所以它们可在以后需要类似数据的查询中通过表缓存来重用，表缓存是一种不需要管理员干预的自动化机制。

高级分析平台

传统上必须将分析工作构建和部署到独立的分析服务器上。这些服务器将运行计算密集型的分析算法并连接后端的一个数据存储库，如数据仓库。此架构延长了从模型开端到部署的时间，需要将数据从数据存储库移到分析服务器。此过程不仅会花太长的时间，而且效率低下，限制了可用于获取洞察的数据，约束了分析建模的范围，还损害了迭代式试验的能力。

Netezza通过IBM Netezza Analytics为结构化数据的高级分析提供了一个与众不同且易于使用的方法，IBM Netezza Analytics是一个高级分析平台，免费包含在每个Netezza设备内。借助IBM Netezza Analytics，可在数据所处的地方(数据仓库内)执行各种分析活动(如数据探索、发现、转换、建模和计分)。这减少了在整个企业中构建和部署分析模型所花的时间。通过缩短从模型开端到部署的时间，公司可在更多的决策中融入富有洞察的按需分析，实现企业级、基于实时的决策。这还让从业者能够试验性地、更快地迭代不同的分析模型，以找到最合适的模型。开发一个模型后，可针对企业中所有相关数据无缝地执行它。预测和计分可在数据所处的地方(数据仓库内)完成。用户可近乎实时地获取预测分数结果，帮助让高级分析便于操作并使其可在整个企业中使用。

IBM Netezza Analytics支持多种工具、语言和框架。它让各种分析应用程序、可视化工具和商业智能工具可通过各种编程方法(如SQL、Java、MapReduce、Python、R、C、C++和Fortran)来驾驭并行化的高级分析；所有这些工具都可用于提供强大、富有洞察的分析。全面的高级分析环境让您可轻松利用此平台，可灵活地使用首选的工具执行临时分析、原型设计和高级分析的生产部署。

Netezza分析平台支持将其强大的内置分析功能集与来自Revolution Analytics(它发布了R的一个商用版本)、SAS、IBM SPSS、Fuzzy Logix和Zementis等供应商的领先分析工具相集成。此外，您可使用该平台的用户定义扩展来开发新功能。完整的Netezza分析平台如图4-5所示。

IBM Netezza Analytics包含一个内置的并行化分析功能库，这个库专为大数据量而设计，让任何分析项目的开发时间曲线变平了。表4-1总结了一些关键的内置分析功能及其优势。

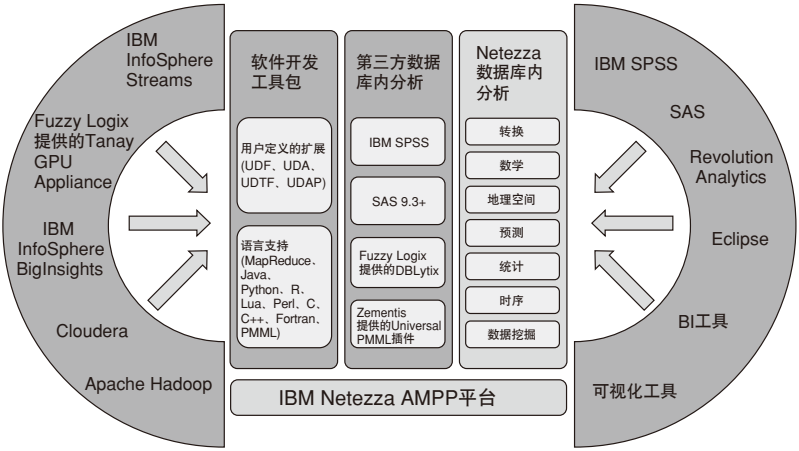


图4-5 IBM Netezza分析平台

转换	执行数据库内数据转换，实现重大的性能提升。
数学	执行深入的数据库内数学计算，以充分利用MPP处理。
统计	计算丰富的统计信息，无需移动数据。
时序	使用丰富的历史数据创建预测并识别趋势，改善模型准确性。
数据挖掘	使用更多数据或所有数据发现新的和新兴的洞察。
预测	从批处理转变为近乎实时的思考速度分析，非常准确且快速地进行预测。
地理空间	在大数据上实现基于位置的分析，提供直接反馈。

表4-1内置于IBM Netezza Analytics中的数据库内分析

使用Hadoop扩展Netezza Analytics平台

如前所述，大规模并行性和靠近源数据处理的原则(体现在IBM大数据平台的所有部分上，包括Netezza、InfoSphere BigInsights [Hadoop]、InfoSphere Streams和InfoSphere Data Explorer)为在大数据集上运行的高级分析提供了明显的好处。因为Hadoop和Netezza都处理已存储的数据，所以我们想解释一下何时使用每种技术，以及两种技术如何协同工作。Hadoop在商用服务器上运行、存储广泛的数据类型、通过MapReduce处理分析查询，并且可预测地随数据量的增长而扩展的能力，使它成为大数据分析一个非常富有吸引力的解决方案。Netezza能将复杂的非SQL算法嵌入到其MPP流的处理元素中，而不会涉及到Hadoop编程的典型细节，这支持低延迟地访问可与广泛的企业BI和ETL工具集成的海量结构化数据。这些原则让Netezza成为了数据仓库融合和高级分析的理想平台。为了充分利用两种技术的优势，Netezza使用一个Hadoop集群提供了不同的连接解决方案(Hadoop将在第5章介绍)。

因为IBM大数据平台非常灵活，所以我们认为值得讨论一下一些结合使用Hadoop和Netezza的典型场景。

探索性分析

有时组织会遇到一个需要分析的新数据源。他们可能对这个新数据源的格式、它包含的数据类型或它封装的关系知之甚少，甚至完全不知道。例如，假设市场部门启动了一项新的多渠道服务，希望将来自Facebook和Twitter的响应与他们可能拥有的其他数据源相集成。如果从未使用过Facebook或Twitter API，或者不熟悉它们的数据源结构，可能需要一定的试验(数据发现)来确定从这个源中提取何种数据，以及如何将它与其他数据来源相集成。Hadoop能够处理还没有为其定义一种模式的数据源，这对此场景非常有用。所以如果希望探索数据内的关系，尤其是在模式不断演化的环境中，Hadoop提供了一种机制来探索数据，直到定义一种正式、可重复的ETL流程。定义该流程并建立结构后，可将数据加载到Netezza中用于标准化的报告或临时分析。

Hadoop就是一卷新磁带: 可查询的归档

大数据分析带来大量数据用于调查分析。我们常常发现，此数据的很大一部分可能时常没有用。这些数据可能是过去的数据或非常细粒度的数据，并且已汇总到数据仓库中。

将所有这些数据放在一个主要针对性能而优化的基础架构中，这可能在经济上不可行。出于此原因，企业可能希望优化他们的分析足迹，将不常访问的数据存储在针对每TB存储的价格进行了优化的基础架构上，并根据需要充分利用更高性能的基础架构。

因为Hadoop的容错分布式存储系统在商用的硬件上运行，所以它可用作此类数据的存储库。不同于没有计算能力的基于磁带的存储系统，Hadoop提供了一种机制来访问和分析数据。因为转移计算比转移数据要更便宜，所以Hadoop的架构更适合用作大数据的可查询归档。事实上，无数的企业客户向我们讲述得最多的用例涉及到Netezza和Hadoop之间的“热-冷”数据存储模式的架构，其中最常使用的数据存储在Netezza中，其他所有数据归档在Hadoop中。当然，除了构建到Hadoop中的本机连接器，IBM提供了一个完整的信息集成和治理平台(将在第10和11章介绍)，可帮助简化并控制此过程。

非结构化数据的分析

关系型数据仓库提供了存储复杂数据类型和非结构化数据的有限能力。而且，通过SQL在非结构化数据上执行计算可能很麻烦且受到局限。Hadoop存储任何格式的数据并使用一种过程编程模式(如MapReduce)分析它的能力使其非常适合存储、管理和处理非结构化数据。您可使用Hadoop预先处理非结构化数据，提取关键特性和元数据，然后将该数据加载到Netezza数据仓库中进行进一步的分析。

客户成功案例: Netezza体验

事实上，数百家组织已经将旧的数据库技术替换为IBM Netezza的数据仓库和分析设备，摆脱了第一代数据仓库带来的挫折感。被替换的系统会强迫用户处理太多的复杂性；仓库需要经过大量培训的专家提供不断的关注和管理。这种复杂性具有双重的危害: 随着数据量增长，管理成本会呈螺旋式上升并失去控制，而离数据很远的业务部门必须寻找专业技术知识才能管理他们与信息的交互。

如前所述，Netezza在数据仓库市场中引发了一场变革，让传统的数据仓库提供商不得不改变他们的战略。为什么在Netezza迅速占领市场时，市场会改变它的进程或加速计划的执行？答案很简单: 客户能够将自己的时间从管理数据的技术方面解放出来，并将这些时间用到业务上以及数据在其成功中可发挥更大作用的方面。以前将多达90%的时间花在低级技术工作上的人现在很少与该技术进行交互，但常常与业务和业务数据交互来研究从数据中创造价值的新方式。我们觉得值得与您分享一些客户案例，这些案例展示了Netezza开启数据仓库设备先河的原因和方式。

T-Mobile: 在PB级数据上轻松提供卓越的性能

每一天，T-Mobile都会处理超过170亿个事件，包括电话呼叫、文本消息及其网络上的数据流量。这会产生高达2PB的需处理数据。T-Mobile需要一个大数据解决方案，它要能够存储和分析多年的呼叫细节记录(CDR)，包括其数百万用户的交换机、帐单和网络事件数据。T-Mobile希望识别和解决网络瓶颈，确保在需要的时间和地点配备所需的质量和容量。

事实证明，Netezza是T-Mobile管理其数据大规模增长的正确解决方案。1,200多位用户正在访问他们的Netezza系统。他们每天会分析超过170亿个事件，执行网络体验质量(QoE)分析、流量规划、流失分析、挂掉的电话分析，以及语音和数据会话分析。自部署Netezza以来，与之前的解决方案相比，T-Mobile显著减少了数据仓库管理活动，还能够使用更大量的细粒度数据来防止虚假索赔，减少税收和呼叫转接费用。更重要的是，T-Mobile能够提高网络可用性，识别和修复所出现的网络瓶颈和拥塞问题。

纽约州立大学: 使用分析帮助查找多发性硬化症的疗法

位于巴法罗的纽约州立大学(SUNY)是一家全球著名的多发性硬化症(MS)研究中心所在地。MS是一种致命的慢性神经疾病，影响着全球近100万人。SUNY团队已经分析从扫描的MS患者基因组获取的数据，以识别其变异可能导致发展为MS的风险的基因。他们的研究人员假定，环境因素和基因因素确定了一个人的MS风险概况。

SUNY的目标是分析临床和患者数据，通过分析性别、地理位置、种族特点、饮食、运动、日照以及生活和工作条件等因素来发现MS患者中的隐藏趋势。他们用于此分析的数据源包括医疗记录、实验室结果、MRI扫描和患者调查；简言之，这是海量的、种类繁多的数据。此类多变量研究中使用的数据集非常大，分析对计算能力的要求极高，因为研究人员正在寻找数千个基因和环境因素之间重要的交互效应。基因环境交互分析中的计算挑战源于一种称为组合爆炸的现象。为了让您对此场景中的数据挖掘所必需的计算量有一定的认识，可考虑1后面有18个0(我们说的是10的18次方)!

SUNY研究人员希望不仅看到哪个变量对MS的发展至关重要，还要看到哪些变量的组合至关重要。他们决定使用Netezza作为数据平台。通过结合使用Revolution R Enterprise for Netezza和IBM Netezza Analytics系统，他们能够使用丰富的变量类型快速构建模型并对涵盖2,000多个可能影响MS的基因和环境因素的庞大数据集运行这些模型。此解决方案帮助SUNY研究人员将所有报告和分析整合到单个位置，

以改善其研究的效率、完备性和影响，将执行此分析所需的时间从27小时缩短到了11分钟！

NYSE Euronext: 减少数据延迟并实现快速的临时搜索

NYSE Euronex是一家欧美公司，经营着多个证券交易所，最著名的是纽约证券交易所(NYSE)和泛欧证券交易所。NYSE每天会获取大约80亿次交易。在某些天，例如在金融市场发生“闪电崩盘”时，他们会处理超过150亿次交易。这相当于每天需要将2TB到4TB数据摄取到其数据仓库中。他们的分析师跟踪一家上市公司的价值，执行趋势分析并搜索欺诈性活动的证据。NYSE和泛欧证券交易所执行市场监督并分析一个交易日的每次交易，这需要对大量数据执行全表扫描。与许多企业的情形一样，NYSE和泛欧证券交易所的传统数据仓库会在存储系统与他们的分析引擎之间来回移动数据，用超过26小时的时间来完成某些类型的处理工作。他们还拥有全球客户，需要无任何中断的24/7访问。他们如何应对这些挑战？他们选择了Netezza。Netezza在数据附近运行计算的核心原则最大限度减少了网络流量，将访问业务关键型数据所需的时间从26小时缩短到了2分钟。它方便了对PB级数据进行快速临时搜索，并实现了全新的分析功能。

5. IBM的企业Hadoop: InfoSphere BigInsights

很少有技术在过去几年获得的关注比Hadoop和NoSQL的还多。再联系到大数据的发展，您有足够的理由编写一图书馆的时尚技术图书。人们对这些技术感到兴奋是有原因的。在处理大数据时，传统的数据存储和分析工具并不得心应手。从数据量的角度讲，超过了数十TB的阈值后，许多工具就开始变得不切实际了。有时“不切实际”指的是该技术无法再进一步扩展，或者它在网上传输用于处理的数据集所花的时间上已达到引爆点。而在其他时候，不切实际指的是尽管技术可以扩展，但处理越来越高的数据量所导致的许可、管理和硬件成本已变得让人无法接受。从种类的角度讲，传统的分析工具仅适用于结构化数据，这些数据最多仅占如今世界所有数据的20%。最后，还有一个数据多快到达组织的问题——大数据速度——下一章会详细介绍。

考虑到人们对可克服静止数据的数据量和种类挑战的技术有着迫切的需求，商业杂志和在线技术论坛热烈地讨论Hadoop和NoSQL就不足为怪了。而且他们并不只是说说而已。大多数财富500强公司中的IT部门都已完成了某种程度的Hadoop等产品的试验。

问题在于，许多这类计划已停滞在“科学项目”阶段。挑战是相同的: 将数据转储到这些存储库中很简单且令人激动，但困难的部分是接下来做什么。对存储在Hadoop中的数据进行有意义的分析需要高度专业化的编程技能——而且对于许多算法，将它们并行化以便在Hadoop中运行非常困难。那么信息治理方面又如何呢？如安全和数据生命周期管理，Hadoop等新技术在这一领域还没有完整的案例。

IBM看到了Hadoop等技术带来的巨大业务变革潜力。这正是IBM拥有大量研究人员、开发人员和支持人员为Hadoop搭建一个平台(称为IBM InfoSphere BigInsights (BigInsights))的原因。BigInsights于2010年10月发布，它具有相对简单的目标: 让Hadoop适合企业使用。本章将介绍企业就绪性的3个主要方面:

- **分析支持** 让不同类别的分析师能够从存储在BigInsights中的数据获取价值，而不要求他们不公正地满足Hadoop编程人员的需求，或者购买昂贵的咨询时间来获得组织中没有人理解的定制Hadoop应用程序。想想可消费性!
- **数据集成** 如果您的企业正在分析存储在Hadoop中的数据，就需要确保您的

Hadoop系统与IT基础架构的剩余部分相集成。例如，您需要能够从数据仓库环境中查询Hadoop数据，以及反过来从Hadoop环境中查询数据仓库数据。

- **卓越运营** 开始依赖于Hadoop生成的分析之后，您需要治理和管理存储在Hadoop中的数据。BigInsights包含许多关键的安全、治理、管理和性能特性。

介绍BigInsights对Hadoop的扩展之前，首先要理解Hadoop是什么。

Hadoop是什么!

非常笼统地讲，Hadoop是一个分布式文件系统和数据处理引擎，专为处理极大量的任何结构的数据而设计。简单来讲，想象一下您有数十台或者甚至数百台(或数千台！)计算机堆叠并通过网络连接在一起。每台计算机(在Hadoop语言中常常称为一个节点)拥有自己的处理器和数十个2TB或3TB的硬盘驱动器。所有这些节点都运行着一个软件，将它们统一到单个集群中，在这个集群中您看到的不是各个计算机，而是一个存储数据的超大卷。这个Hadoop系统的美妙之处在于，您可将任何数据存储在此空间中：数百万张抵押合同扫描图像、数天或数周的安全摄像机片段、数万亿条传感器生成的日志记录，或者来自一个呼叫中心的所有操作员文字备注。这种无需担忧数据模型的数据获取方法实际上是NoSQL运动的一个重要原则(称为“延迟模式(schema later)”)。相对而言，传统的SQL和关系型数据库领域依赖于相反的方法(“即时模式(schema now)”)，其中数据模型是获取数据时最重要的关注点。在这方面，Hadoop的灵活性更加显著。它不仅是您转储许多文件的地方。您还可以在基于Hadoop的数据库中存储各种不同的模型：关系、列和键/值。换句话说，Hadoop中既有完全非结构化的数据，也有完全关系化的数据，以及介于它们之间任何一点的数据。我们这里介绍的数据存储系统称为Hadoop分布式文件系统(HDFS)。

我们回头看看这个包含许多单独节点的虚构Hadoop集群。假设您的企业使用此集群存储电子商务网站的所有单击流日志记录。您的Hadoop集群使用BigInsights发行版，您和分析师决定对此数据运行一些会话式分析，以隔离留下购物车而去的顾客的常见模式——我们将此用例称为最后一英里优化。运行此应用程序时，Hadoop将应用程序逻辑的副本发送给集群中的每个计算机，使用每个计算机本地的数据运行该程序。所以无需将数据移动到中央计算机进行处理(将数据带给功能)，而是应用程序移动到存储数据的所有位置(将功能带给数据)。这种编程模型称为Hadoop MapReduce。

大象的起源：Hadoop发展历史

现在您对Hadoop的概念已有了大体的认识，接下来分析一下它为何导致了IT革命。首先看看Hadoop的起源。

Hadoop的灵感来源于Google针对其Google(分布式)文件系统(GFS)和MapReduce编程模式的研究工作。IBM长期参与MapReduce的研究，他从2007年10月就开始与Google合作，针对大规模Internet问题联合执行一些与MapReduce和GFS相关的大学研究。在Google(分别于2003和2004年)发布描述GFS和MapReduce的文章后不久，开源社区(由Doug Cutting领导)的人们就将这些工具应用到了开源Nutch搜索引擎中。人们很快就发现，分布式文件系统和MapReduce模块组合的应用范围远不止是搜索。2006年早期，这些软件组件变成了他们自己的研究项目，称为Hadoop。Hadoop是一个非常怪异的名称(您还会在Hadoop领域发现许多怪异的名称)。阅读当今的任何Hadoop图书，它们很可能最初都将该名称用作其项目的吉祥物，所以我们也这么做。Hadoop实际上是创建者Doug Cutting的儿子为他的填充玩具大象取的名字。在为他的项目构想名称时，Cutting显然想找个朗朗上口又没有任何具体意义的东西，所以他儿子的玩具的名称似乎很完美。

构建Hadoop的许多工作是由Yahoo!完成的，而Hadoop的许多灵感也来源于搜索引擎行业，这并不是巧合。随着Internet范围的数据处理在大型集中化服务器上变得愈加不切实际，Hadoop对Yahoo!和Google都是一次不可或缺的创新。唯一的替代方案是横向扩展，将存储和处理工作分散到集群中的数千个节点上。据报告，Yahoo!拥有40,000多个节点分散在它的Hadoop集群中，这些集群存储了超过40PB的数据。

Hadoop开源团队从Google借鉴了这些概念，让它们适用于更广泛的用例。不同于事务性系统，Hadoop旨在通过一个高度可扩展、分布式的批处理系统来扫描大数据集并生成结果。

Hadoop不关乎像思考那么快的响应时间、实时仓库或极高的事务速度，它关乎发现，以及从可伸缩性和分析角度将以前的几乎不可能变成可能。

Hadoop的组件和相关项目

前面已提到，Hadoop一般被视为有两个部分：一个文件系统(HDFS)和一个编程范型(MapReduce)。Hadoop的一个重要组件是内置到环境中的冗余性。数据不仅冗余地存储在整个集群的多个位置中，编程模型也是如此，所以故障会被预料到并通过运行集群中各种服务器上的程序部分来自动加以解决。得益于这种冗余性，可以将数据和编程工作分散在一个非常大的商用组件集群中，就像之前讨论的集群一样。众所周知，商用硬件组件会发生故障(尤其是在拥有非常多的商用硬件组件时)，但这种冗余性提供了容错能力和Hadoop集群的自我修复能力。这让Hadoop能够将工作负载分散在大型的廉价机器集群中，以解决大数据问题。

市面上有许多Hadoop相关项目，一些较著名的项目包括: Apache Avro(针对数据序列化)、Cassandra和HBase(数据库)、Hive(提供类似SQL的临时查询来执行数据聚合和汇总)、Mahout(一个机器学习库)、Pig(一种高级Hadoop编程语言，为并行计算提供了一种数据流语言和执行框架)，以及ZooKeeper(为分布式应用程序提供协调服务)。由于本书的篇幅所限，我们不会介绍这些相关项目，但网络上有许多信息，当然BigDataUniversity.com上也有。

Hadoop 2.0

只要Hadoop是一个围绕IT的流行谈话主题，就会不可避免地提及NameNode单点故障(SPOF)。有趣的是，对于有关具体设计限制的谈论，记录在案的NameNode故障实际上非常少，这是对HDFS恢复能力的有力证明。但对于任务关键型应用程序，甚至一次故障也是多余的，所以为NameNode提供大量的备份对企业更广泛地采用Hadoop极其重要。MapReduce处理层(JobTracker服务器)也是一个单点故障位置。

在Hadoop 1.0部署中，您可通过主动/被动故障转移解决方案来解决NameNode可用性问题。一个选项是在两个独立的Hadoop集群之间自动复制数据。另一个选项是为主要节点(其中包含NameNode，也可包含JobTracker服务)提供一个专用备份，以便Hadoop集群的所有NameNode元数据都得以备份。在一个NameNode发生故障时，Hadoop集群可使用备份的NameNode重新启动。

在Hadoop 2.0中(在编写本文时还处于alpha状态)，有两处重要的可用性改进: 为HDFS NameNode指定一个热备份的能力和YARN(也称为MapReduce2)，后者可分配JobTracker功能，让此服务器不再是一个SPOF。IBM提交者正在与开源社区携手合作，让Hadoop 2.0适合生产，当然，IBM打算借助最新的企业就绪开源创新来保持BigInsights的先进性。

在开源社区中，一些供应商已在其产品的正式上市(GA)版本中发布了Hadoop 2.0代码，这引发了一些公开争论。争论的重点在于，这里包含的Hadoop 2.0代码未分类为生产就绪。其他供应商(包括IBM和Hortonworks)已避开了这种方法，重申他们的策略是在其发行版中仅发布生产就绪的开源代码。

我们希望进入更深的层面，介绍HDFS和MapReduce的工作原理，更不用说Apache Hadoop生态系统中的其他项目，但篇幅有限。要想更详细地了解Hadoop，请访问BigDataUniversity.com，您在那里可找到许多免费、高质量的在线课程!

内部组成: InfoSphere BigInsights的组件

在介绍BigInsights的企业就绪性功能之前，我们希望建立一个背景，从许可、封装以及软件组件的角度介绍产品的主要组件。我们将介绍BigInsights中包含的Hadoop组件、BigInsights的主要用户界面(BigInsights Web控制台)、开发人员工具，以及各种可用的版本。

InfoSphere BigInsights 2.0中包含的Hadoop组件

BigInsights包含Apache Hadoop及其相关开源项目作为一个核心组件。这个组件被通俗地称为IBM Distribution for Hadoop。IBM仍然致力于这些开源项目的完整性，将确保与它们100%兼容。这种对开源的忠实性提供了许多好处。对于为其他与开源100%兼容的发行版开发代码的人，他们的应用程序也将在BigInsights上运行，反之亦然。这种开源兼容性让IBM积累了100多个BigInsights合作伙伴，包括数十家软件供应商。简单来讲，如果软件供应商为开源Hadoop使用了库和接口，它们也将适用于BigInsights。

IBM还在发布BigInsights的定期产品更新，以便它具有开源组件的最新版本。

下表列出了BigInsights 2.0中包含的开源项目(和它们的版本)，在编写本文时BigInsights 2.0是最新版。

组件	版本
Hadoop(常见实用程序、HDFS和MapReduce框架)	1.0.3
Avro(数据序列化)	1.6.3
Chukwa(监视大型集群化系统)	0.5.0
Flume(数据收集和聚合)	0.9.4
HBase(实时读取和写入数据库)	0.94.0
HCatalog(表和存储管理)	0.4.0
Hive(数据汇总和查询)	0.9.0
Lucene(文本搜索)	3.3.0
Oozie(工作流和作业编排)	3.2.0
Pig(编程和查询语言)	0.10.1
Sqoop(Hadoop与数据库之间的数据传输)	1.4.1
ZooKeeper(流程协调)	3.4.3

在每个BigInsights版本中，对开源组件和IBM组件的更新都会经历一系列测试周期，确保它们可协同工作。这是我们希望澄清的另一个特殊之处：您无法将新代码直接放到生产环境中。根据我们的经验，开源项目中始终存在向后兼容性问题。BigInsights消除了与您的Hadoop组件的典型开源项目相关的许多风险和猜测问题。

它会经历与其他IBM软件相同的严格回归测试和质量保证测试流程。所以您可以问自己：您愿意成为自己的系统集成商，重复地测试所有Hadoop组件以确保兼容性吗？或者愿意让IBM找到一个您可安心部署的稳定软件体系吗？

IBM的开源项目支持传统在Hadoop上得到了传承。Big Insights开发团队拥有Hadoop项目和相关项目的一些提交者。这些提交者为开源项目的领袖地位做出了突出贡献，拥有将IBM代码捐赠给开源代码库的使命。通过让提交者加入BigInsights开发团队，IBM更深入地参与到开源Hadoop的持续演化中。这样，IBM还能更快地查找和消除开源代码中的错误。我们与开源Hadoop社区的共同希望是，我们都从UNIX战争中学到了教训，在科技公司构建自己的UNIX版本来解决相同的问题时，他们浪费了数十年的开发工作。借助开源Hadoop，即使参与的公司是竞争对手，他们也具有让Hadoop变得更好这一共同兴趣。

BigInsights Web控制台

BigInsights带给Hadoop的一个出色功能是称为BigInsights Web控制台的富界面(参见图5-1)。无论您在世界另一边一个集群中的一个公共云实例上运行，还是在公司服务器场中的一个1,000节点集群上运行，这个控制台都是整个集群的焦点，因为所有管理、应用程序部署和应用程序执行活动都在这里执行。BigInsights Web控制台拥有许多功能，我们将在后面几节介绍它们，但现在请注意图5-1左侧的任务帮助，它们专为提高适用性而设计。

您在控制台中能看到的内容取决于您作为BigInsights用户的访问级别。例如，如果拥有管理员帐户，您可看到管理仪表板，如Application Status和Cluster Status。如果拥有用户帐户，您只能看到适用于浏览文件、运行应用程序和执行分析工作的仪表板，也就是File Explorer、Application Dashboard和BigSheets界面。

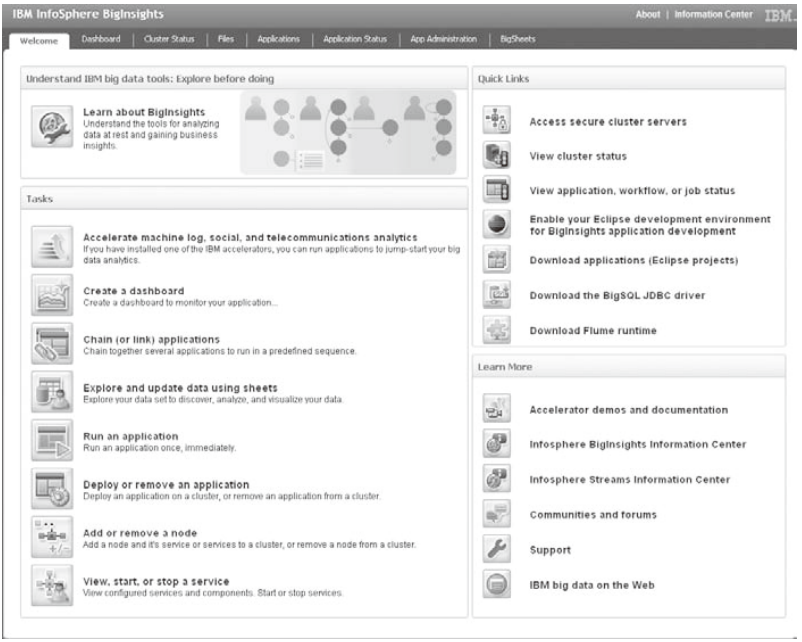


图5-1 BigInsights Web控制台

BigInsights开发工具

除了Web控制台，BigInsights还包含一组Eclipse插件，用于开发可处理大数据的应用程序。这个插件包可从Web控制台下载，包含配套的安装说明。这些开发工具包含两个Eclipse透视图: BigInsights(适用于使用SQL、Jaql Hive、HBase、Pig或MapReduce开发数据处理应用程序)；以及BigInsights Text Analytics(指导您执行一个构建文本分析应用程序的工作流)。大数据平台还包含其他透视图，如用于构建IBM InfoSphere Streams应用程序的透视图。

BigInsights开发工具可连接到Web控制台来了解集群。这让开发人员能够轻松地测试和部署应用程序，直接处理集群。在后面介绍分析的一节中，我们将介绍BigInsights Web控制台和开发工具如何协同工作，提供构建、测试和部署应用程序的完整生命周期。

BigInsights版本: 基础版和高级版

BigInsights有两个版本: 一个免费产品(基础版)和一个付费产品(企业版)。BigInsights基础版是IBM集成、测试和预先配置的一

个下载版本，适用于想体验Hadoop的任何人。这个版本包含BigInsights安装程序，让您快速地、只需进行单击即可实现自己的Hadoop集群。作为一个开源组件包，这代表着IBM Distribution for Hadoop。除了一些数据库连接器和一些工具功能，这个发行版中未包含任何分析、集成或卓越运营特性。如果希望升级到企业版，这可以无缝地完成。

BigInsights企业版包含BigInsights基础版中的所有特性，以及本章后面将介绍的分析、集成和卓越运营功能。此许可中还包含Streams的有限使用许可，它允许您使用BigInsights结合运行各种流数据作业和静止数据处理(这里有许多集成，下一章将详细介绍)。为了反映不同类型的部署，BigInsights企业版也有一个非生产许可和一个入门许可。

部署BigInsights

Hadoop(以及相关的BigInsights)设计为在许多具有专用存储和处理资源的单独计算机组成的一个集群上部署。Hadoop中内置了故障转移和冗余功能，但它可部署在廉价甚至不可靠的硬件上。我们将此类计算资源称为商用硬件，具体来讲，如具有标准的旋转式磁盘存储的Intel芯片组。这与我们看到部署在数据仓库服务器中的健全、高质量，而且肯定很昂贵的硬件形成了鲜明的对比。

易用性: 简单的安装流程

BigInsights安装程序的设计非常简单。IBM的开发团队问他们自己，“IBM如何能减少让Hadoop正常运行所需的时间，而无需正常情况下让开源软件正常运行所需的工作和技术技能？”他们通过BigInsights回答了这个问题。

BigInsights安装程序的主要目标是让您远离复杂性。这样，您无需担忧必备的软件或确定下载哪些Apache Hadoop组件。您无需担忧这些组件之间的配置或Hadoop集群的总体设置。BigInsights安装程序会为您完成所有这些工作，您需要的只是单击一个按钮。Hadoop启动复杂性被BigInsights完全消除了。简单来讲，您的体验与任何商用软件的安装将非常类似。

而且，您可使用IBM的安装程序图形化地构建一个响应文件，然后可使用该文件以自动化的方式将BigInsights部署到集群中的所有节点上。

一种低成本的入门方式: 在云上运行BigInsights

随着虚拟化技术的出现，云已成为企业利用IT资源的一个流行平台。鉴于采购和构建一个适当规模的Hadoop集群要涉及大量的工作，似乎云技术的设计考虑到了Hadoop。

我们假设您想知道BigInsights可为您的企业

做什么。您拥有一个大数据集，拥有分析师和开发人员在准备试验它。如果考虑执行现场试验性的Hadoop安装，常规来讲采购过程和最终安装硬件可能要花几星期(乃至几个月)。另一方面，您可以在几小时内让一个定做的基于云的BigInsights集群正常运行。我们将此场景稍微延伸一下，想象您的数据加载到这个基于云的集群上，而且您的分析师在夜以继日地工作。如果某些方面的性能很差，没有关系！可动态地向集群添加更多资源(内存、数据节点、CPU)。这是云成为试用部署的一个优秀选择的另一个原因：无需花费大量时间掌握参考架构的细节。有了云，如果部署无法工作，可实时进行调整。

像IBM大数据平台中的其他组件一样，BigInsights可很好地融入云部署中。从技术角度讲，BigInsights支持虚拟化，这是云环境中的必备要素。BigInsights还拥有卓越的安全性，整个集群可以隔离在单个访问点之后，只能通过Web控制台访问。从许可角度讲，再简单不过了，BigInsights的定价模型基于(在复制之前和压缩之后)集群中存储的数据量。最后。BigInsights许可对虚拟化没有任何限制。再次说明，您只需为所使用的资源付费，这是与云服务相同的定价模型！

BigInsights已在一些最著名的云提供商中成功部署：如Amazon EC2、RackSpace和IBM SmartCloud。IBM还与RightScale(领先的云管理平台)进行合作。这种合作让IBM支持在一些公共云和私有云上试用BigInsights以及其他一些IBM产品，这些云构建于来自Eucalyptus和Cloud.com的开源云体系之上。

更高等级的硬件：IBM PowerLinux Solution for Big Data

尽管目前为止商用硬件部署已成为惯例，但我们也看到了一些替代方案。一个有吸引力的选项是IBM PowerLinux Solution for Big Data。这是一个针对Hadoop优化的Linux on Power硬件解决方案。

Power硬件的一个最重要的优势就是可靠性，因此非常适合将其部署为集群中的主要节点。如果需要更高的能效或需要让DataNode专门处理较少的内容，或者只需要更多的处理能力，Power绝对是最佳的途径。

2011年，IBM Watson系统击败了美国智力竞赛节目Jeopardy!上的两位最优秀的冠军。Linux on Power是Watson的底层平台，它也充分利用了Hadoop来处理它

的一些子系统。在开发IBM PowerLinux Solution for Big Data期间，各种定制功能以及从Watson项目学到的经验都应用到了这个产品中。

Cloudera支持

在BigInsights 1.4发布后，IBM扩展了它的大数据平台，使之能在其他Hadoop发行版上运行，首先是Cloudera。Cloudera品牌已变得与Hadoop联系紧密，这主要得益于他们的Cloudera Distribution for Hadoop (CDH)的广泛应用。尽管BigInsights包含IBM Distribution for Hadoop，但对CDH的扩展支持并没有反映出一种冲突性的方法。反之，它是IBM支持一个强大且统一的Hadoop社区的另一种方式。因为它完全基于开源代码，所以IBM Distribution for Hadoop和CDH的实际差别不是很大。

IBM对Cloudera(以及最终的其他开源Hadoop发行版)支持的主要目的是强调IBM大数据平台的价值。简而言之，即使您不使用IBM Distribution for Hadoop，也可利用BigInsights中的高级分析、集成和工具功能。客户可从Cloudera获取CDH，安装他们的CDH集群，然后可将BigInsights企业特性部署到CDH之上。对于CDH用户，这是一个好消息: 他们不再需要艰难地选择离开自己舒适的安装内容，而去使用BigInsights。这里的重点在于，选择使用哪个发行版没有体系中更高层级的分析方法重要。

分析: 探索、开发和部署

采用Hadoop的组织所面临的最大挑战是，如何快速且轻松地从其集群中存储的数据获取价值。这里需要考虑许多因素，一些因素已在前面提到。例如: 并行化众多分析算法的困难性；让存储在Hadoop中的数据可供业务分析师和数据科学家访问；以及处理凌乱的数据。BigInsights通过在其各种组件上应用一种集成的且基于平台的方法，解决了所有这些问题。

作为Hadoop的分析平台，BigInsights支持3类分析用户: 业务线分析师、数据科学家和应用程序开发人员。此外，BigInsights的分析和开发组件具有生命周期管理工具，这让开发人员能够轻松地将分析应用程序部署到业务线用户和数据科学家。图5-2显示了BigInsights中的所有分析组件。本节将介绍其中每个组件。我们还会介绍应用程序生命周期方法和BigInsights组件背后的相应特性。

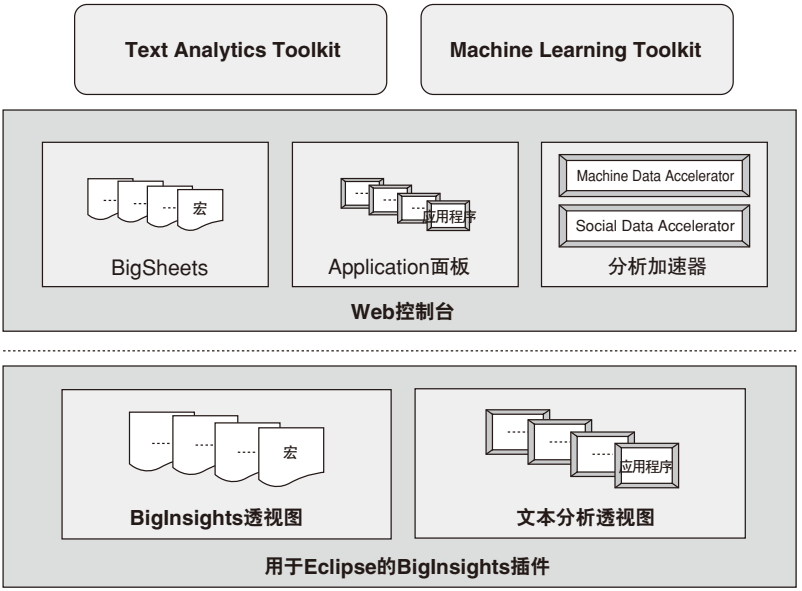


图5-2 BigInsights中的数据分析组件

Advanced Text Analytics Toolkit

我们最喜欢的一个包含IBM大数据平台的深度集成示例是Advanced Text Analytics Toolkit。尽管BigInsights分析组件都连接到此工具包，但它也集成到Streams产品中。这意味着您为组织的数据编写的文本提取器可(通过BigInsights)部署在静止数据上或(通过Streams)部署在运动数据上。Text Analytics Toolkit也集成在所有BigInsights组件中。本章后面的"BigInsights应用程序生命周期"一节将介绍，您能部署文本提取器，从而在Web控制台中轻松地执行操作。因为Advanced Text Analytics Toolkit不仅仅是一项BigInsights特性，我们会用专门的一章来介绍它——请参阅第8章了解详细说明。

面向大众的机器学习：BigInsights上深入的统计分析

BigInsights包含一个Machine Learning Toolkit，它为统计员和数学家提供了一个平台来对BigInsights集群中的数据执行高性能的统计和预测分析。它包含一种高级机器学习语言，该语言在语义上类似于R(用于统计计算的开源语言)，分析师可使用它为其数据处理工作应用统计模型。该工具包中

还包含丰富的内置数据挖掘算法和统计模型。Machine Data Accelerator和Social Data Accelerator(下一节将介绍)都使用了这些统计算法来执行报告和可视化工作。BigInsights Web控制台的Applications面板中以应用程序的形式提供了一些机器学习算法和实用程序。其中包括一些数据处理实用程序，例如将分隔开的数字数据转换为矩阵格式，以及线性回归和多级支持矢量机(support vector machines，SVM)等算法。

Machine Learning Toolkit包含一个引擎，它将以机器学习语言形式表达的统计工作负载转换为并行化的MapReduce代码，这意味着它向分析师隐藏了这一复杂性，让分析师可以专心完成本职工作。简而言之，分析师不需要成为Java编程人员，他们不需要在分析应用程序中考虑MapReduce。

Machine Learning Toolkit是IBM Research一个性能专家、统计学家和数学家团队开发的。他们的主要目标是为需要在Hadoop上下文中执行复杂统计分析的分析师提供高性能和易用性。因此，此工具包中包含了用于生成低级MapReduce执行计划的优化技术。因此与在MapReduce或其他Hadoop语言(如Pig)中直接实现的算法相比，统计作业能够实现几个数量级的性能提升。不仅分析师可在他们的分析应用程序中避免使用MapReduce编码技术，他们编写的机器学习代码也针对卓越的Hadoop性能进行了高度优化。简单来讲，很难编写并行程序——IBM提供了一种声明性语言来构建并行驱动的统计应用程序，这种方式与它发明SQL时为关系型数据库使用的方式相同。

BigInsights还集成了R，它在BigInsights Web控制台的Applications面板中以一个应用程序的形式提供。借助此应用程序，用户可对其数据运行临时的R脚本。

分析加速器: 大海捞针?

认识到可反映具体用例的工具的重要性，IBM构建了一组大数据分析加速器，这些工具旨在让分析师和数据科学家能够富有成效地探索数据。最初的加速器集合包括: SocialData Accelerator、Telecommunications Data Accelerator和Machine Data Accelerator。这些加速器已经过配置，可提供开箱即用的价值且可进一步定制。例如，您可针对公司的日志数据来调整Machine Data Accelerator的文本分析提取器。请参阅第9章，了解这些分析加速器的详细说明。

适合大众的应用程序: 定制应用程序的轻松部署和执行

BigInsights Web控制台有一个称为Applications的选项卡，您可在其中看到许多随时都可运行的应用程序。IBM依照如今在移动设备领域很常见的“应用程序商店”

的概念来设计此面板。运行这些应用程序只需一种简单、通用的方法: 单击应用程序，输入要从哪个路径读取需分析的数据，输入另一个希望将结果集存储到哪个路径(以及应用程序的任何其他字段)，您就可以运行该应用程序了。无需将宝贵的预算花在使用安装程序部署应用程序等操作上，只需单击即可运行。

BigInsights附带了许多应用程序，您可使用它们测试集群，了解您可构建的应用程序类型。实际上，您可以用Eclipse BigInsights项目的形式从控制台下载BigInsights随带的每个应用程序的代码，所以您自己的应用程序项目已有了一个现成的起点！此特性的另一个好处在于，您可使用它作为一个前端调度程序，供最终用户运行那些分散在企业中的大数据应用程序。

业务分析师和数据科学家可利用现有的应用程序作为构建块来创建新应用程序。为此，用户可使用一个图形化的拖放工具将一个工作流中的现有应用程序链接起来。您可轻松地将一个应用程序的输出提供给该链中的下一个应用程序作为输入。

BigInsights对安全性的重视从这里可见一斑，因为具有应用程序管理员角色的用户可确定将授权哪些用户运行具体的应用程序。如果数据源或服务需要您提供安全凭据，应用程序接口支持您利用BigInsights凭据存储，让您能够安全地将身份验证信息传递给您连接到的数据源或服务。

数据发现和可视化: BigSheets

尽管Hadoop使分析大数据成为可能，但您可能需要成为一位编程人员，对探索数据的MapReduce模式具有出色的理解。我们已看到在尝试向非技术人员解释并行编程概念时发生的情况，显然大多数业务分析师对理解Hadoop中存储的数据具有很大的障碍。BigInsights提供了答案: 一个名为BigSheets的基于浏览器的可视化工具。此工具让业务线用户能够使用熟悉的电子表格界面驾驭Hadoop的力量。BigSheets不需要任何编程操作(它在幕后自动生成Hadoop代码)或特殊的管理。如果您会使用电子表格(数据透视表、切片、切块等)，就可使用BigSheets对任何结构的海量数据执行分析。

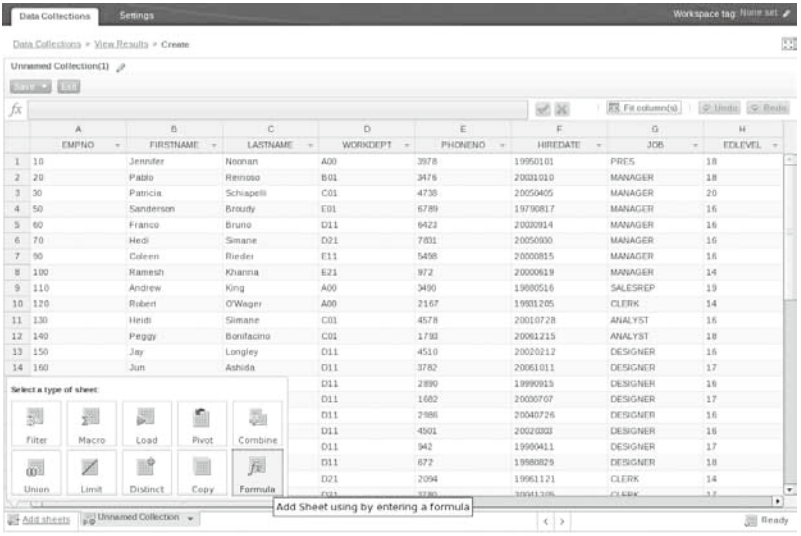
BigSheets执行大数据分析涉及到3个简单步骤:

1. 收集数据。您可从多个来源收集数据，使用在BigInsights中部署的应用程序爬取网页、本地文件或您网络上的文件。支持多种协议和格式，包括HTTP、HDFS、Amazon S3 Native File System(s3n)和Amazon S3 Block File System(s3)等。还有一个设施可通过数据导入定制插件来扩展BigSheet。例如，可构建一个插件来获取Twitter数据并将其包含在BigSheets集合中。

2. 提取和分析数据。收集数据后，可在电子表格界面中看到它的一个抽样，如图5-3

所示。现在，您可使用BigSheets中提供的电子表格类型的工具来操作这些数据。例如，可将来自不同集合的列组合在一起，运行公式，或者过滤数据。也可为BigSheets部署定制的宏。例如，可部署一个使用BigInsights开发工具构建的文本提取器作为一个BigSheets宏。构建表格和细化分析时，您可在抽样数据中看到中间结果。只有在单击Run按钮时，才会将分析应用到整个数据集合。因为您的数据可能具有不同的大小，从几GB到TB，再到几PB，所以迭代式地处理一个小数据集是最佳的方法。

3. 探索和可视化数据。通过表格对数据运行分析后，可应用可视化来帮助理解数据。BigSheets提供了一些传统和新兴的大数据可视化工具，包括:



	A	B	C	D	E	F	G	H
	EMPNO	FIRSTNAME	LASTNAME	WORKDEPT	PHONENO	HIREDATE	JOB	EDLEVEL
1	10	Jennifer	Noonan	A00	3978	19950101	PRES	18
2	20	Pablo	Reinoso	B01	3476	20031010	MANAGER	18
3	30	Patricia	Schiapelli	C01	4738	20050405	MANAGER	20
4	50	Sanderson	Broudy	E01	6789	19790817	MANAGER	16
5	60	Franco	Bruno	D11	6423	20030914	MANAGER	16
6	70	Heidi	Simane	D21	7831	20050000	MANAGER	16
7	90	Coleen	Rieder	E11	5498	20000815	MANAGER	16
8	100	Ramesh	Khanna	E21	972	20000619	MANAGER	14
9	110	Andrew	King	A00	3490	19980516	SALESREP	19
10	120	Robert	O'Wager	A00	2167	19931205	CLERK	14
11	130	Heidi	Simane	C01	4578	20010728	ANALYST	16
12	140	Peggy	Bonifacino	C01	1793	20061215	ANALYST	18
13	150	Jay	Longley	D11	4510	20020212	DESIGNER	16
14	160	Jun	Ashida	D11	3782	20061011	DESIGNER	17
				D11	2890	19990915	DESIGNER	16
				D11	1682	20030707	DESIGNER	17
				D11	2986	20040726	DESIGNER	16
				D11	4501	20020303	DESIGNER	16
				D11	942	19990411	DESIGNER	17
				D11	672	19980829	DESIGNER	18
				D21	2094	19961121	CLERK	14

图5-3 BigSheets的电子表格式界面

- **标记云** 显示词频；字母越大，引用该词汇的数据就出现得越多。参见图5-4中的示例。
- **饼图** 显示比例关系，其中各个切片的相对大小表示它所占的数据比例。
- **地图** 显示覆盖到一个世界地图或美国地图上的数据值。
- **热图** 类似于地图，但增加了一个维度来显示覆盖在一个地图上的数据值的相对密度。
- **条形图** 显示一个指定列的值的出现频率。

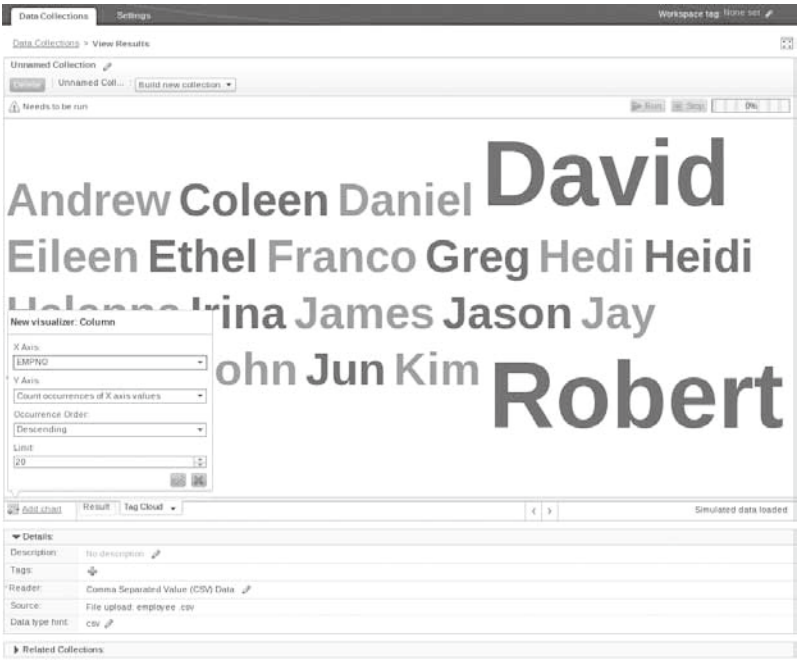


图5-4 BigSheets中的数据可视化

BigSheets的可视化工具具有很高的扩展能力。因此，可包含定制的插件来专业地呈现数据。

BigInsights开发环境

本章前面已简略地提到，BigInsights的一个主要组件是一组Eclipse插件，它们为创建BigInsights应用程序提供了一个功能丰富的开发环境。这包括您通常会与一个集成开发环境(IDE)关联起来的所有可用性特性，如内容完成、上下文敏感的帮助、设计时帮助和错误检测、断点和调试控制台、代码编辑器、工作流助手、测试工具、一个部署工具等。

可使用BigInsights开发环境开发文本提取器(第8章将详细介绍与开发这些工件有关的优势)、大数据查询(包括独立SQL、用于Hive来源的HiveQL和HBase表达式)和大数据应用程序(包括Jaql、Pig和MapReduce)。使用这些语言时，BigInsights开发环境为您提供了更好的体验。在Eclipse编辑器中，可以看到代码突出显示和语法错误检查。还为支持的语言提供了模板和集成的帮助，以帮助您更快入门。

除了BigInsights开发环境中的查询和应用程序开发工具，还有一个图形化的工作流编辑器可用于轻松地开发Oozie工作流(参见图5-5)。如果没有该工具，这将是一个痛苦的过程，涉及到根据Oozie模式手动编辑一个XML文件。使用工作流编辑器，您可拖放Oozie工作流元素(Start、End、Kill、Action、Action、Fork、Join和Decision)。还有一个代码编辑器(包含代码突出显示和语法错误检测功能)，您可在其中编辑已生成的Oozie XML。

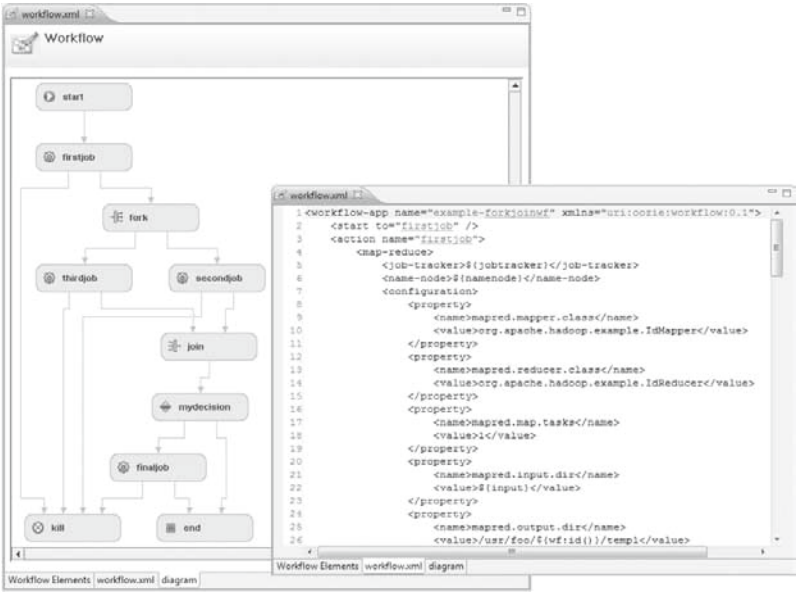


图5-5 应用程序链图形界面

为了加快应用程序部署速度，您可配置BigInsights开发环境感知您的BigInsights集群，这样只需单击一个按钮即可将应用程序代码推送到集群中。

事实是，Hadoop的专业编程人员很少，但知道Eclipse或SQL的开发人员很多。想象一下，从概念证明到企业生产，最初需要少量的开发人员，而最后需要一个完整的开发、架构和质量保证团队——可能有数百名开发人员。BigInsights开发环境的设计宗旨是只需一个普通开发人员的技能集就能使用它。所以如果您的企业开始使用BigInsights，您拥有的集成工具可帮助开发团队更快地上手。

BigInsights应用程序生命周期

现在我们已全面介绍了BigInsights中的分析和应用程序开发工具，让我们回头看看BigInsights中整合的应用程序生命周期管理的全景图。下面详细描述了各个生命周期步骤及其影响的组件(另请参见图5-6)。

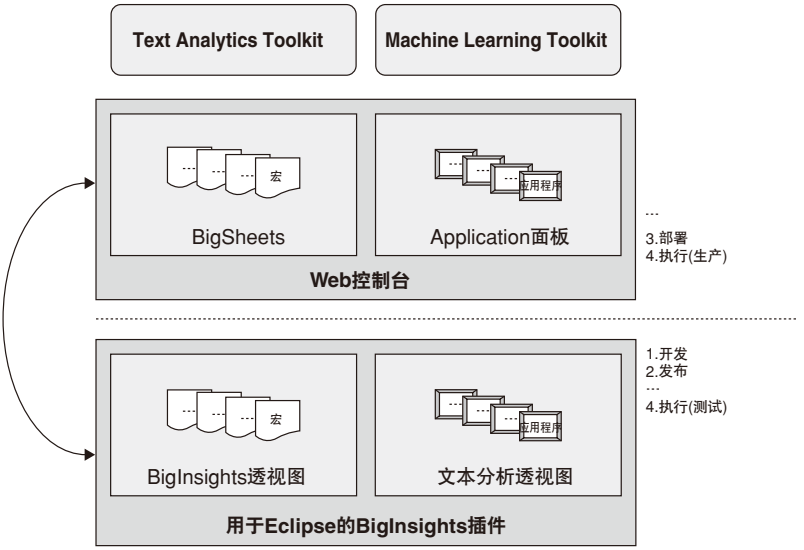


图5-6 BigInsights应用程序生命周期

1. **开发** 使用针对Eclipse的BigInsights开发工具，您可创建文本提取器，将连接故障隔离到您的Web服务器日志中。

2. **发布** 通过Eclipse，您可推出文本提取器，在Web控制台中将其用作应用程序(通过Applications面板)或BigSheets的一个宏。

3. **部署** 通过Web控制台，具有应用程序管理员角色的用户可部署和配置一个应用程序，以在BigInsights集群上执行。

4. **执行** 最终用户可在控制台中运行文本提取器(作为应用程序或宏，具体取决于它的发布方式)。开发人员也可从Web控制台运行应用程序，下载结果并在文本分析调试工具中的测试工作内使用这些结果。

这些是IT组织在Hadoop上下文中开发和部署应用程序需执行的所有步骤。在BigInsights中，IBM提供了广泛的自动化和

编排工具，从而让此流程更加快捷、更加简单和更加可靠。没有此工具，您不仅需要手动移动文件和设置配置，还需要为应用程序开发包装器和接口，让用户可执行它们。

数据集成

IBM大数据愿景的一个关键价值是集成各种来源的数据的重要性。Hadoop不是一个适用于所有情况的解决方案，不能满足您的所有存储和处理需求，所以企业中仍然需要继续使用其他存储库。但您承诺将一个Hadoop集群添加到IT基础架构后，其他一些系统不可避免地就需要访问它。正是考虑到这一点，IBM为BigInsights中的数据开发了一个SQL接口。这消除了需要查询Hadoop数据的DBA的学习曲线。此外，BigInsights支持与许多来源交换数据，包括IBM PureData System for Analytics(以前称为Netezza)、IBM PureData System for Operational Analytics(以前称为IBM Smart Analytics System，由DB2提供支持)；DB2 for Linux, UNIX, and Windows；其他通过Java Database Connectivity(JDBC)接口实现的关系型数据存储；InfoSphere Streams；InfoSphere Information Server(具体讲是 DataStage)等。本节将深入详细地介绍一些集成点，以及简略介绍其他集成点；我们简略介绍的集成点将在本书后面的相应章节中详细介绍。

基于分析的IBM PureData Systems和DB2

可通过两种方式在BigInsights与任何基于分析的IBM PureData Systems(Analytics和Operational Analytics版，或者以前的Netezza或DB2 for Linux, UNIX, and Windows版本)交换数据：从您的数据库服务器，或者从您的BigInsights集群。这些IBM数据库技术都可通过一组可安装在数据库服务器上的用户定义函数(UDF)连接BigInsights并与之进行交互。BigInsights可通过一个特殊的高速适配器连接到IBM PureData System for Analytics和Netezza。BigInsights可通过JDBC模块(后面一节将介绍)连接到IBM PureData System for Operational Analytics或DB2。

BigInsights UDF

与BigInsights、基于分析的IBM PureData Systems、Netezza和DB2的集成有两个主要组件：一组安装在数据库服务器上的UDF，以及BigInsights集群上的一个Jaql服务器(用于监听来自这些UDF的请求)。Jaql服务器是一个中间件组件，可接受由Netezza(运行Netezza Analytics 2.0或更高版本)或DB2服务器(9.5或更高版本)技术提供支持的任何系统的Jaql查询处理请求。具体来讲，Jaql服务器可接受来自一个受支持的数据库DB2服务器的以下类型的Jaql查询：

- 从BigInsights集群读取数据。
- 在BigInsights集群中上传(或删除)Jaql代码模块。
- 提交Jaql作业(它们可能引用您之前从支持的数据库服务器上传的模块)，从而在BigInsights集群上运行该作业。
- (仅支持IBM PureData System for Analytics)从数据仓库将数据写入BigInsights集群。

运行这些BigInsights功能让您能够从传统的应用程序框架轻松地与Hadoop集成。借助这些功能，数据库应用程序(它们在平时无法感知Hadoop)可利用在获取关系型数据时使用的相同SQL接口来访问BigInsights集群中的数据。这些应用程序现在可利用BigInsights集群的并行性和规模性，无需额外的配置或其他开销。尽管与传统的Hadoop应用程序相比，此方法会带来额外的性能开销，但它是将大数据处理集成到现有IT应用程序基础架构中非常有用的方式。

IBM PureData System for Analytics适配器

BigInsights包含的一个连接器支持在BigInsights集群与IBM PureData System for Anlaytics(或它的前身 - Netezza设备)之间交换数据。这个适配器支持拆分表(一种类似拆分文件的概念)。这要求将表分区并将每个已拆分的部分分配给一个特定的映射器。这样就能并行地处理您的SQL语句。

该适配器利用Netezza技术的外部表特性，可将该特性视为一种具体化的外部UNIX管道。外部表使用JDBC。在此场景中，每个映射器充当着一个数据库客户端。基本而言，一个映射器(作为客户端)将连接到数据库，并开始从IBM PureData System的基础架构所创建的UNIX文件读取数据。

JDBC模块

Jaql JDBC模块支持您在任何具有标准JDBC驱动程序的关系型数据库中读取和写入数据。这意味着您可使用当今市场中的每种主流数据库仓库产品轻松地交换数据和发出SQL语句。

借助Jaql的MapReduce集成。每个映射任务可访问一个表的一个特定部分，这样能够针对已分区的数据库并行处理SQL语句。

用于运动数据的InfoSphere Streams

您将在第7章中发现，Streams是IBM的流数据实时分析解决方案。Streams包含一个BigInsights接收适配器，它允许您将流数据直接存储到BigInsights集群中。Streams还包含一个用于BigInsights的来源适配器，它允许Streams应用程序从集群中读取数据。BigInsights与Streams之间的集成带来了一些有趣的可能性。从总体上讲，您可创建一个基础架构来实时响应在数据中检测到的更改(因为数据由Streams动态处理)，同时使

用丰富的现有数据(由BigInsights静态地存储和分析)来提供响应。也可使用Streams作为一个大规模数据摄取引擎，以过滤、装饰或操作一个要存储在BigInsights集群中的数据流。

使用BigInsights接收适配器，Streams应用程序可为BigInsights集群编写一个控制文件。可配置BigInsights来响应这类文件的出现，以便它触发一个要在集群中运行的深入分析操作。对于更高级的场景，来自Streams的触发器文件也可包含查询参数，以定制BigInsights中的分析。

InfoSphere DataStage

DataStage是一个提取、转换和加载(ETL)平台，能够集成来自大量异构来源和目标系统的海量数据。它提供了一个易于使用的界面和设计工具；支持大规模扩展；可转换多个不同信息来源的数据；以及批量或实时提供数据。通过扩展它作为数据集成代理的角色，DataStage已得到扩展，能够应用于BigInsights且可向BigInsights集群推送数据和从中拉取数据。DataStage与BigInsights的连接器集成了HDFS，可充分利用集群化的架构，所以任何对同一个文件的批量写入工作都可并行完成。DataStage集成的结果是，BigInsights可迅速与任何能够连接到DataStage的软件产品交换数据。

我们越来越多地看到Hadoop被用作一种动态ETL引擎，尤其是用于非结构化数据。但是，ETL并不只是转换。流程编排、调试、生命周期管理和数据血统都是一些重要的考虑因素。DataStage和InfoSphere Information Server平台的剩余部分为处理所有这些任务提供了途径。甚至已有计划在Information Server与BigInsights之间建立更紧密的连接，例如从DataStage设计BigInsights作业的能力，它们使强大且灵活的数据转换场景变为可能。第10和11章将介绍Hadoop优化，包括DataStage的Big Data File Stage (BDFS)。

卓越运营

随着越来越多的组织依靠Hadoop来执行业务分析，对更多治理和管理功能的需求不断增多。人们逐渐习惯于企业关系型数据库中丰富的数据管理特性，希望在Hadoop集群中也能看到它们。与此同时，Hadoop正进入一个人们对数据治理的理解已高度发展的时代。关系型数据库的优势在于它在发展过程中融入了许多数据理念。结果，许多经验丰富的IT人员对Hadoop具有很高的期望。麻烦在于，开源软件更多地专注于核心功能，而不是扩展一些管理性特性。在IBM，有数百位研究人员和开发人员，他们都是治理、工作负载管理和绩效优化方面的行业领先专家。从Apache Hadoop项目启动以来，许多专家都参与了已合并到BigInsights中的Hadoop解决方案的开发工作中。

保护集群的安全

安全是企业软件的一个重要问题，而对于开源的Hadoop，在投入生产之前有一些限制需要考虑。好消息是，BigInsights通过保护对管理性界面的访问、关键的Hadoop服务、锁定开放端口、基于角色的安全、与InfoSphere Guardium Database Security (Guardium)的集成等减少了安全攻击面，从而解决了这些问题。

BigInsights Web控制台被设计为集群的一个网关。它具有更高的安全性，支持轻量型目录访问协议(LDAP)身份验证。LDAP和反向代理支持可帮助管理员将访问限制到授权的用户。此外，集群外部的客户端必须使用受保护的REST接口才能通过该网关访问集群。相对而言，Apache Hadoop在集群中的每个节点上都拥有开放的端口。您需要打开的端口越多(开源Hadoop中有许多这样的端口)，环境就越不安全，因此无法通过内部审计扫描的可能性就越高。

可配置BigInsights与一个LDAP凭据服务器通信，从而执行身份验证工作。控制台与LDAP服务器之间的所有通信使用LDAP(默认)或同时使用LDAP和LDAPS(基于HTTPS的LDAP)来执行。BigInsights安装程序可帮助您定义LDAP用户和组与4个BigInsights角色(系统管理员、数据管理员、应用程序管理员和用户)之间的对应关系。安装BigInsights后，您可在LDAP中添加或删除用户，以授予或撤销对各种控制台功能的访问权。

Kerberos安全已集成到开源Hadoop中，它提供了服务和一些操作工具，但不支持备用的身份验证方法。BigInsights使用LDAP作为默认的身份验证协议。BigInsights重视LDAP的使用，因为相对于Kerberos和其他替代方案，它是一个更容易安装和配置的协议。最后，BigInsights支持备用的身份验证选项，如Linux Pluggable Authentication Modules(PAM)。可使用此选项部署Kerberos令牌身份验证，甚至是计量生物学身份验证。

将集群放在Web控制台的软件防火墙背后并建立用户角色，这有助于锁定BigInsights及其数据，但一个完整的安全方案必须包含制度合规性。例如，任何接受信用卡的企业都必须遵守支付卡行业数据安全标准(PCI DSS)，这需要保护客户数据并记录任何访问。Guardium是合规性市场的领导者，它可处理关系型数据库生成的审计日志。IBM构建了BigInsights的扩展，已存储对集群中的数据进行访问的审计日志。任何涉及到以下组件的数据访问活动都会生成审计日志: Hadoop RPC、HDFS、MapReduce、Oozie和HBase。

Apache Hadoop的最新更改提高了HDFS安全性；但是，仍然有一些限制。在这方面，IBM拥有的企业锁定经验(已整合到

BigInsights中)让您能够打造更安全、强大和更容易维护的多租户解决方案。

监视集群的所有方面

为了帮助您管理集群，BigInsights Web控制台提供了集群各个组件的一个实时、交互式的视图。BigInsights控制台包含检查BigInsights环境(包括集群中的节点、Hadoop和BigInsights服务、HDFS元数据、MapReduce度量指标和您的作业[应用程序]的状态)健康状况的图形化工具。

监视集群

BigInsights拥有一个功能丰富的监视仪表板，支持您调整管理性指标的查看方式。以下区域具有预定义的仪表板: 系统(包含CPU利用率和平均负载)、集群、HDFS(包含总文件数和损坏的数据块)和MapReduce(包含正在运行的Map和Reduce任务，以及已失败的任务)。每个仪表板包含多个监视小部件，您可非常详细地配置它们，包括从时间范围到刷新率(参见图5-7)。

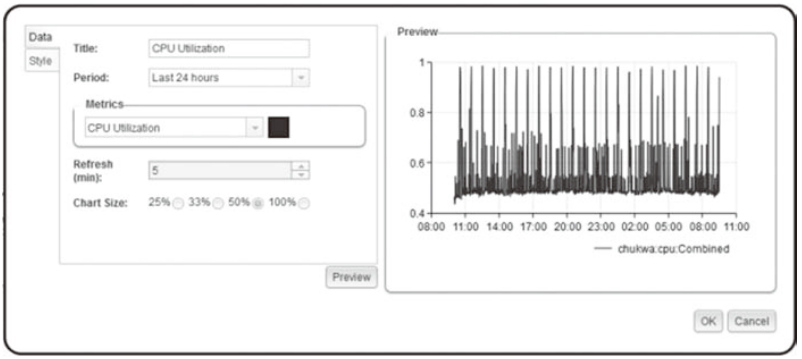


图5-7 定制各种监视小部件

监视应用程序

BigInsights Web控制台提供了集群的上下文敏感视图，所以人们只能根据其关联的角色看到所需的内容。如果有个具有安全角色“用户”的人运行一个程序，那么他只能在Application History视图中看到自己的统计数据。这个特定的视图进行了优化，能显示高级的应用程序状态信息，隐藏低级的工作流和作业信息。

拥有管理员角色的人可在Application History视图和Application Status监视仪表板中看到所有应用程序的状态，并且能够进入到各个工作流和作业以执行调试或性能测试工作。Application Status窗格拥有针对工

作流和作业的视图。每个视图列出活动的和已完成的工作流和作业。您可深入到每个工作流或作业来了解进一步的细节，也可从这里查看相关的元素。

压缩

处理Hadoop环境中预计会出现的大量数据时，压缩的想法很有吸引力。一方面，您可节省大量空间(尤其是考虑到每个存储块默认都会在Hadoop中复制3次时)；另一方面，由于传输的数据量更少而提高了数据传输速度。在选择一种压缩模式之前，应考虑两个重要的因素：您使用的压缩算法的可拆分压缩以及压缩和解压速度。

可拆分压缩

在Hadoop中，如果文件大于集群的块大小设置(通常每个块对应一个拆分部分)，文件会被拆分。对于未压缩的文件，这意味着各个文件拆分部分可由不同的映射器并行处理。图5-8显示了一个具有垂直行的未压缩文件，每行表示拆分部分和块的边界(在本例中，拆分部分与块具有相同的大小)。

Big data represents	a new era in data	exploration and	utilization, and IBM	is uniquely positioned	to help clients design,	develop and execute	a big data strategy.

图5-8 Hadoop中一个未压缩的可拆分文件

如果文件(尤其是文本文件)已压缩，麻烦就会出现。对于大多数压缩算法，无法将各个文件拆分部分与同一个文件的其他拆分部分分开解压。更具体地说，这些压缩算法是不可拆分的(在探讨压缩和Hadoop时请记住这个关键词语)。在最新的Hadoop生产版本中(编写本文时为1.0.3)，不支持拆分已压缩的文本文件。对于应用了Sequence或Avro格式的文件，这不是问题，因为这些格式内置了同步点，因此是可拆分的。对于不可拆分的压缩文本文件，MapReduce处理仅限于单个映射器。

例如，假设图5-8中的文件是您的Hadoop集群中的一个1GB的文本文件，并且您的块大小设置为BigInsights默认的128MB，这意味着您的文件涵盖8个块。使用Hadoop中提供的传统算法压缩此文件时，无法并行处理每个已压缩文件的拆分部分，因为该文件只能作为整体来解压，而不能基于各个拆分部分来解压各个部分。图5-9描绘了这个文件的压缩(和二进制)状态，其中不可能单独解压各个拆分部分。注意到不匹配的地方了吗？(拆分部分的边界为虚线，块边界为实线。)

因为Hadoop 1.0.3没有为可拆分的文本压缩提供原生支持，所以已压缩的文本文件的所有拆分部分只能由单个映射器来处理。对于许多工作负载，这会导致显著的性能下降，从而让压缩成为一个不可行的选项。但是，可配置Jaql理解文本文件的可拆分压缩，并自动使用并行映射器处理它们。对于其他环境，您可手动这么做(如Pig和MapReduce程序)，使用TextInputFormat输入格式代替Hadoop标准。

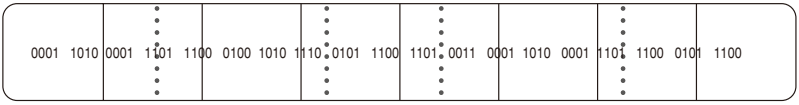


图 5-9一个已压缩的不可拆分文件

压缩和解压速度

俗语“天下没有免费的午餐”在压缩上得到了切实的体现。根本没有什么魔法；从本质上讲，您只是消耗了CPU周期来节省磁盘空间。所以，让我们做一个这样的假设：压缩Hadoop集群中的数据会降低性能，因为将数据写入集群时，压缩算法(它们是CPU密集型的)需要CPU周期和时间来压缩数据。类似地，在读取数据时，任何针对已压缩数据执行的MapReduce工作负载都会影响性能，因为需要CPU周期和时间来解压已压缩的数据。这带来了一个难题：您需要平衡存储节省和额外的性能开销之间的优先级。

我们应注意到，如果一个应用程序具有I/O限制(许多仓库风格的应用程序通常是这样)，您可能会看到应用程序的性能提升，因为具有I/O限制的系统通常具有备用的CPU周期(在CPU中表现为空闲的I/O等待)可用于运行压缩和解压算法。例如，如果使用空闲的I/O等待CPU周期来执行压缩工作，而且具有很好的压缩率，那么最终会有更多的数据流经I/O管道，这意味着这些应用程序需要更快的性能来从磁盘抓取大量数据。

一项BigInsights优势：IBM CMX压缩

BigInsights包含IBM CMX压缩(LZO压缩编解码器的一个IBM版本)，它支持拆分已压缩的文件并让各个已压缩内容的拆分部分可由MapReduce作业并行处理。

一些Hadoop在线论坛描述了如何使用LZO的GNU版本来实现可拆分的压缩，那么为什么IBM要创建自己的版本呢，为什么不使用GNU LZO代替？首先，IBM CMX压缩编解码器不会在压缩文件时创建索引，因为它使用固定长度的压缩块。相反，GNU LZO算法使用可变长度的压缩快，这就需要一个

索引文件来告诉映射器它可在何处安全地拆分一个已压缩文件。(对于GNU LZO压缩，这意味着映射器需要在解压和读取操作期间执行索引查找。此索引会导致管理性开销，因为如果移动已压缩的文件，就需要移动相应的索引文件)。第二，许多公司(包括IBM)拥有法律政策，禁止他们购买或发布包含GNU Public License(GPL)组件的软件。这意味着在线Hadoop论坛中描述的方法需要额外的管理性开销和配置工作。此外，一些企业的政策还限制部署GPL代码。IBM CMX压缩与BigInsights全面集成，遵循与BigInsights的剩余部分相同的企业友好的许可协议，这意味着您可更轻松地使用它，没有与GPL替代方案相关的麻烦。

在未来的Hadoop版本中，bzip2算法将支持拆分。但是，bzip2的解压速度比IBM CMX慢得多，所以bzip2不是注重性能的工作负载的理想压缩算法选择。

图5-10显示了之前示例中的已压缩文本文件，但它处于一种可拆分状态，其中各个拆分部分可由它们自己的映射器解压。注意，拆分部分的大小相同，表示这是固定长度的压缩块。

在下表中，您可看到BigInsights平台上提供的4种压缩算法(IBM CMX、bzip2、gzip和DEFLATE)及其一些特征。

压缩编解码器	文件扩展名	是否可拆分	压缩程度	解压速度
IBM CMX	• cmx	是	中	最快
bzip2	• bz2	是，但还不可用	最高	慢
Gzip	• gz	否	高	快
DEFLATE	• deflate	否	高	快

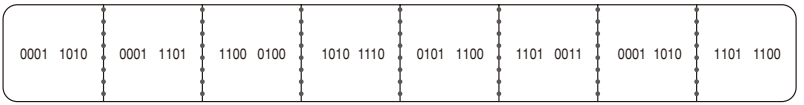


图5-10 一个可拆分的已压缩文本文件

更好的工作负载调度: Intelligent Scheduler

开源Hadoop随带了一个基本的先进先出(FIFO)调度程序，以及一个支持备用调度选项的可插拔架构。您可通过Apache Hadoop项目获得两个可插拔的调度工具:

Fair Scheduler和Capacity Scheduler。这些调度程序很相似，它们都只需极少的资源就能处理较小的作业，以避免资源枯竭。这些调度程序没有提供足够的控件来确保实现最优的集群性能，或者为管理员提供实现可定制的工作负载管理所需的灵活性。例如，Fair非常擅长确保将资源应用到了工作负载上，但它没有为您提供类似SLA的细粒度控制。

可以想象，IBM在工作负载管理方面拥有数十年的经验和研究专长。从大型机到分布式系统，IBM在不断反复证明着这些专业技能。例如，IBM是我们知道将(DB2中的)数据库级工作负载管理与底层主机操作系统的工作负载管理功能相集成(以便它们能联合工作)的唯一一家数据库供应商。数据库中的数据或文件系统中的数据的工作负载管理仍然属于工作负载管理范畴。

IBM Research的性能专家已在研究Hadoop中的工作负载调度问题，并且已设计了一个名为Intelligent Scheduler(以前称为FLEX Scheduler)的解决方案。这个调度程序扩展了Fair Scheduler，通过持续调整分配给作业的最小插槽数量来操作它。Intelligent Scheduler包含各种不同的度量指标，您可使用它们优化工作负载。这些度量指标可由管理员在集群级别上进行选择，或者由个人在作业级别上进行选择。也可选择为这些度量指标分配权重来平衡竞争优先级，最小化所有作业度量指标的总和，或者最大化它们的总和。

以下是您可用于调优工作负载的一些Intelligent Scheduler设置示例:

- **平均响应时间** 调度程序将最多的资源分配给较小的作业，确保这些作业能快速完成。
- **最大拉伸性** 调度程序按作业所需资源量的比例来分配资源。大型作业拥有更高的优先级。
- **用户优先级** 调度程序将最多的资源分配给一个特定用户的作业。

自适应MapReduce

在IBM Research进一步优化Hadoop性能的工作中，他们开发了一个名为“自适应MapReduce”的概念，这个概念扩展了Hadoop，让各个映射器能感知自己的情况并感知其他映射器的情况。这种方法让各个映射任务能适应其环境并制定高效的决策。(如果考虑要为数据库可伸缩性实现哪种连接池和连接密度，并将其应用到Hadoop的自适应MapReduce，就有了一个很好的类比入手，如果您还不熟悉此领域的话。)

在正常的MapReduce中，当一个作业即将开始时，Hadoop会将数据拆分为许多片断，这些片断称为拆分部分。每个拆分部分都分配有一个映射器。要确保一个均衡的工作负载，这些映射器呈波浪状部署，而且新映射器在旧映射器完成其拆分部分的处理后才会启动。在此模型中，较小的拆分部分大

小意味着映射器更多，这有助于确保实现均衡的工作负载和最大限度降低故障成本。但是，由于每个映射任务具有更高的启动成本，较小的拆分部分也会导致集群的开销增加。对于具有更高的映射任务启动成本的工作负载，更大的拆分大小可能更高效。运行映射任务的自适应方法让BigInsights同时具备了两个领域的优势。

自适应MapReduce的一个组件是自适应映射器的概念。自适应映射器扩展了传统的Hadoop映射器的概念，

方法是跟踪一个中央存储库中的文件各个拆分部分的状态。每一次一个自适应映射器完成一个拆分部分的处理，它就会咨询这个中央存储库，锁定另一个拆分部分以供处理，直到作业完成。这意味着对于自适应映射器，只会部署一批映射器，因为各个映射器仍然可使用其他拆分部分。锁定一个新拆分部分的性能成本比一个新映射器的启动成本要少得多，这相当于显著提高了性能。图5-11的左侧显示了一个集相似度连接(set-similarity join)工作负载的基准测试结果，结果表明使用自适应映射器减轻了较高的映射任务的启动成本。自适应映射器结果(参见AM栏)基于32MB的较小拆分部分大小。仅使用了一批映射器，所以通过避免更多映射器的启动成本而显著提高了性能。

对于某些工作负载，任何均衡性的缺失都可能被更大的拆分部分大小放大，这会导致其他性能问题。使用自适应映射器时，可(没有损失地)避免不均衡的工作负载，方法是调优作业，从而使用更小的拆分部分大小。因为只有一批映射器可用，所以您的工作负载不会因许多额外的映射器启动成本而陷入瘫痪。图5-11的右侧显示了TERASORT记录上一个连接查询的基准测试结果，其中在个别映射任务之间产生了不均衡性，进而导致较大的拆分部分大小出现不均衡的工作负载。自适应映射器结果(参见AM栏)基于32MB的较小拆分部分大小。仅使用了一批映射器，所以通过避免额外的映射器启动成本而实现了显著的性能提升。

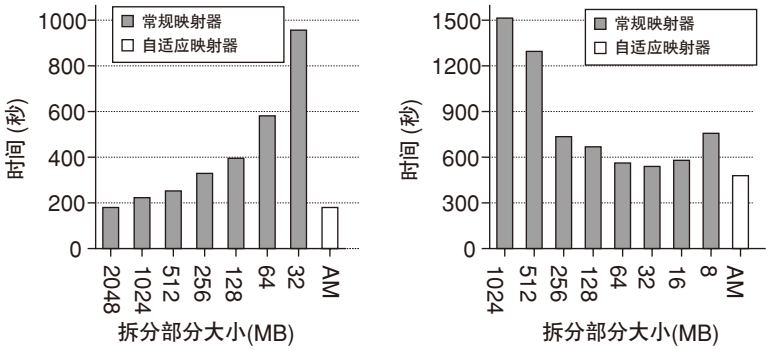


图5-11 对一个集相似度连接工作负载的基准测试，其中通过使用自适应映射器减少了较高映射任务的启动成本

灵活的Hadoop文件系统: GPFS-FPO

General Parallel File System (GPFS)由IBM Research在上世纪90年代为高性能计算(HPC)应用而开发。自1998年第一次发布以来，全球许多最快的超级计算机上都使用了GPFS，包括Blue Gene、Watson(Jeopardy!超级计算机)以及当今全球最快的超级计算机IBM Sequoia。除了HPC，GPFS也经常在全球数千个其他任务关键型安装中找到。毋庸置疑，GPFS因极高的可伸缩性、高性能和可靠性而赢得了企业级的声誉和血统。

HDFS背后的设计原则由一些用例定义，这些用例假设Hadoop工作负载将涉及到非常大的文件集的顺序读取(集群中还不存在随机文件写入——只有追加写入)。相对而言，GPFS是为众多不同的工作负载和大量用途而设计的。

针对Hadoop扩展GPFS: GPFS File Placement Optimization

2009年，IBM扩展了GPFS，使其适用于Hadoop。GPFS-Shared Nothing Cluster最终被重命名为GPFS-FPO(File Placement Optimization)。GPFS最初只能用作存储区域网络(SAN)文件系统，它通常不适合Hadoop集群，因为在存储数据的节点与处理它们的节点相同(这需要数据位置感知性)时，MapReduce作业的性能会更高。在SAN中，数据的位置是透明的，这需要很高的网络带宽和磁盘I/O，尤其是在拥有许多节点的集群中。

通过将Hadoop数据存储在GPFS-FPO中，您就摆脱了HDFS中固有的基于设计的限

制。您可充分利用GPFS-FPO的血统作为Hadoop集群中一个多用途文件系统，这会带来极高的灵活性。

GPFS-FPO与HDFS之间的一个重大的架构区别是，GPFS-FPO是一个内核级文件系统，而HDFS在操作系统之上运行。HDFS中的许多限制源于它无法全面兼容POSIX的事实。而另一方面，GPFS-FPO完全兼容POSIX，这带来了极高的灵活性。

因为GPFS兼容POSIX，所以存储在GPFS-FPO中的文件可被所有应用程序看到，就像存储在计算机中的任何其他文件一样。例如，在复制文件时，任何授权用户都可使用传统操作系统命令来列出、复制和移动GPFS-FPO中的文件。而这在HDFS中行不通，用户需要登录Hadoop才能看到集群中的文件。对于复制或备份，HDFS唯一可用的机制是通过Hadoop命令shell手动复制文件。

GPFS-FPO全面兼容POSIX让您能够轻松管理Hadoop存储，就像IT环境中的任何其他计算机一样。这在打造Hadoop技能时可为您带来规模经济效益，让生活更加轻松。例如，传统的文件管理实用程序可以工作，就像您的备份和还原工具和过程一样。GPFS-FPO将实际扩展您的备份功能，因为它包含时间点(PiT)快照备份、非现场复制和其他实用程序。

借助GPFS-FPO，您可为Hadoop集群使用一个强大的关注点分离基础架构，安全地管理多租户Hadoop集群，让其他应用程序能够共享集群资源。这在HDFS中是不可能的。从容量规划角度将这也有所帮助，因为如果没有GPFS-FPO，您就需要提前设计专门分配给Hadoop集群的磁盘空间。事实是，您不仅要评估需在HDFS中存储多少数据，还要猜测为MapReduce作业的输出提供多少存储空间，这可能因工作负载的不同而差别巨大。最后，不要忘记您还需要考虑Hadoop系统创建的日志文件将占用的空间！借助GPFS-FPO，只需要关心磁盘本身的装载状况，无需为Hadoop专门分配存储空间。

让GPFS成为大规模任务关键型IT安装的首选文件系统的所有特征，也适用于GPFS-FPO。毕竟，它仍然是GPFS，但拥有Hadoop友好的扩展。您会在GPFS-FPO中获得相同的稳定性、灵活性和性能，以及过去习惯使用的所有实用程序。GPFS-FPO还提供了分层存储管理(HSM)功能，它可高效地管理和使用具有不同检索速度的磁盘驱动器。这让您能够管理多温度数据，将“热”数据放在性能最佳的硬件上。HDFS可不具备此能力。

GPFS-FPO改变了游戏规则，它于2010年赢得了权威的超级计算存储挑战赛大奖，被评为提交给这次竞赛的“最富有创新的存储解决方案”。

注意 GPFS-FPO目前仅用于beta测试。

小结

本书前面已经提到，真正需要理解的是大数据并不等同于Hadoop。Hadoop只是您解决如今各种挑战(以及您还不知道的挑战)所需的多个数据处理引擎之一。出于此原因，IBM长期致力于Hadoop研究，为Apache项目做出了卓越贡献并继续打造这个生态系统。借助其在企业级基础架构和优化方面的悠久历史，IBM通过BigInsights将此经验应用到Hadoop之上。BigInsights包含开源的Hadoop并添加了一些卓越运营特性，如大数据优化的压缩、工作负载管理、调度功能，甚至是应用程序开发和部署生态系统。

尽管卓越运营带来了规模经济效益和人们对基础架构信任，但Hadoop真正的潜力在于分析功能——更准确地讲，分析功能的实用性。出于此原因，BigInsights包含了许多工具包、加速器和工具，如BigSheets，他所追求的是让人们普遍使用大数据，让大数据不会牢牢禁锢在需要大批Java编程人员来挖掘数据价值的高深技术部门中。此外，BigInsights提供了一个端到端文本提取框架。简单地说，Hadoop没有提供这些功能，所以IBM构建了一个高价值的产品体系，将Hadoop视为一类居民并集成在整个IBM大数据平台中。

我们认为，要展示如今Hadoop集群所需的任何类型的分析价值，非IBM对Cloudera的Hadoop发行版中BigInsights内的各种分析功能的支持莫属。这可总结为一句话: 它展示了您需要与其一起踏上大数据旅程的合作伙伴类型: 可帮助您从数据中获取价值的合作伙伴。毕竟，我们还未看到任何人会对无法提供任何价值的100节点集群投以尊重的目光。

分析运动中的大数据

6 利用InfoSphere Streams进行实时分析处理

为什么我们愿意等待可操作的洞察？我们如何学会了要忍受数天、数周或数月才可以识别新的机会、了解我们的业务状况，或识别顾客对我们产品和服务的顾虑？随着世界的发展速度越来越快，继续等待简直是不可接受的。正如IBM利用InfoSphere BigInsights (BigInsights)和IBM PureData System for Analytics(在2012年第四季度宣布，是替代Netezza的新IBM品牌)提供了独特地处理静止数据最大分析问题的技术，IBM还提供了对运动数据进行分析，从而即时从数据获得价值的能力。涉及到导航如今企业可用的信息海洋时，使用BigInsights和基于分析的任何IBM PureData Systems可以为您提供多种功能，而IBM InfoSphere Streams (Streams)则可从流经企业的数据之河中提供大量的洞察。您可以选择发掘这些河流，及时地为您的业务获得竞争优势，您也可以站在一旁，看着机会白白溜走。这就是使用Streams的原因。其设计让您可以利用大规模并行处理(MPP)技术在数据流过时对其进行分析，这样您就可以实时了解发生了什么，并有选择地采取行动，制定更好的决策，并改善结果。

本书到了这里，我们可以确定，您已经了解大型数据库速度特征的重要性，以及通过运动平台将静止分析的集成移到您的业务前沿如何能真正提高您的大数据IQ。出于这个原因，本章会比该书的其他章节更详细地从技术角度进行介绍。但是，先澄清我们所说的Streams和流(stream)的意义：Streams是指IBM InfoSphere Streams产品，流指的是数据流。了解这一点后，让我们看看Streams的基础知识，这是定义其工作方式以及用例的技术基础。

基础知识: InfoSphere Streams

Streams是一个强大的分析计算软件平台，在内存中的数据被存储到磁盘之前对这些数据进行持续分析和转换。它并没有收集大量数据，而是处理数据并将数据存储在磁盘上，然后对其进行分析，与其他分析方法一样，Streams让您可以直接对运动中的数据应用分析。使用Streams对运动中的数据进行分析时，您能够以最快速度获得结果，并有可能获得显著的硬件节省以及最高的吞吐量。

既然有这么多优势，您可能会问：“有什么收获？”为了实现这些优势，Streams主要对在整个集群的内存中所维护的数据“窗口”进行操作。然而，大内存容量让数据窗口可以代表几秒钟到几天的数据分析，具体取决于数据流速率。可利用在处理过程中所累积的上下文与来自静止引擎(如Hadoop)或数据库的数据来丰富此数据。

我们一般建议在以下用例中使用Streams:

- 实时识别事件，如确定在社交媒体上的客户情绪何时变得更加负面
- 关联与结合在时间上密切相关的事件，如在日志文件中的警告，随后是一个系统中断
- 持续计算组合的汇总，如股市中每个行业内每个股票的价格趋势

对于需要在一个大型数据集执行多次传递，以便对回归模型进行评分的分析，我们一般建议不使用Streams，而应该使用BigInsights或一个专用的IBM分析数据仓库引擎对静止数据进行分析。然而，通过使用由BigInsights、IBM PureData System for Operational Analytics(之前被称为IBM Smart Analytics System)、IBM PureData System for Analytics或其他分析工具(如SPSS)构建的模型，Streams可以充分利用丰富的历史上下文。

如果您已经很熟悉Complex Event Processing(CEP，复杂事件处理)系统，可能会在Streams中看到一些相似的地方。然而，Streams的设计更具扩展性和动态性，以支持更为复杂的分析，并且比其他系统支持更高的数据流速率。许多CEP或流处理系统(包括Storm等新的开源项目)在整个集群中每秒通告几十万个事件。相比之下，IBM Streams技术已被证明可以在单个服务器上每秒处理几百万个事件——它很快。(不要忘了，您能够以接近线性的可扩展性在集群中部署Streams)。

此外，Streams具有更好的企业级特征，如高可用性、一个功能丰富且易于使用的应用程序开发工具集、无数开箱即用的分析功能，以及与多个常见企业系统的集成。事实上，Streams在其标准工具包中提供了近30个内置的运算符，在扩展工具包中提供数十个运算符(如数据挖掘和文本分析)和数百个函数，以方便应用程序的开发。甚至还有一个CEP工具包，其中提供了通常可以在CEP系统中找到的Streams函数。虽然其他系统无法与强大的Streams相比，但CEP系统和开源项目(如Storm)的出现凸显了对运动数据进行分析的重要性正在日益增加。

您可以将一个Streams应用程序视为一组互联的运算符。可将多个运算符组合成单个可配置的部署单元，该部署单元称为一个作业，一个应用程序是由一个或多个作业组成的。通常在取消前会一直部署应用程

序，连续处理通过它的数据流。将数据带进Streams应用程序的运算符通常称为源适配器。这些运算符读取输入流，然后产生一个可供下游运算符使用的流。分析步骤包括多个运算符，它们根据一个或多个流的输入来执行特定的操作。最后，对于进入持续分析平台的各种方式都有多种方式输出，在Streams中，这些输出被称为接收适配器(sink adapter)。(想像水从水龙头流出，并流进厨房的水槽)。任何运算符都可以同时是源适配器、标准运算符和接收适配器，但将它们认为是不同的运算符将会有所帮助。我们在本章的后面将介绍所有这些运算符。

Streams应用程序一般分为以下两种类型。第一种无论发生什么事情都保持最新数据，即使这意味着放弃旧数据也是如此。对于这些应用程序，现在比处理每一位数据更重要。需要保持最新数据的应用程序示例包括检测和响应网络攻击、在股市上做出买卖决定，或监测一个人的生命体征。对于这些应用程序，您需要根据最新的数据制定决策。Streams通过提供可扩展性和高吞吐量、负载分流(load-shedding)运算符，以在必要时智能地减少数据，并维护系统级的高可用性，以保持应用程序的运行，从而支持这种类型的应用程序。

第二种类型的应用程序要求始终处理数据的每一个位。这些应用程序在过去通常使用数据库技术，但出于效率、及时性，或无法跟上数据速度等原因，部分应用程序被迁移到Streams。这种类型的应用程序示例是处理电信行业的呼叫详细记录(CDR)。出于合规性和业务方面的原因，需要对这些CDR删除重复记录、进行转换，并拼接在一起，但不能丢失任何一个记录。通过CDR仍然在运动中时将处理工作迁移到Streams，可以实现显著的效率提高，并带来全新的业务机会。对于这些类型的应用程序，固有的高吞吐量加上系统级的高可用性和应用程序模式，让Streams成为一个很好的选择。

与Web应用服务器或数据库一样，您可以定制Streams，以交付能够解决业务问题的应用程序。Streams提供了许多企业级特性，如在数十或数百台服务器的集群中部署应用程序的能力，以提高可用性、可扩展性和性能(类似于Hadoop，但针对实时应用程序)。Streams还提供了一个易于使用的拖放式工具环境，可帮助您设计并构建流式应用程序(在本章的“Streams编程一点通”一节中进行介绍)。另一个不错的特性是，Streams与BigInsights共享相同的Text Analytics Toolkit，让您能够在整个大数据平台中重用各种技能和代码片段(我们会在第8章中讨论这一点)。

您可以为单个服务器或集群构建应用程序，Streams自动将运算符融合成高度优化的处理元素(PE)，这些元素在集群中以点对点的方式传输流式数据，以实现最佳的吞吐量。

准备好部署流式应用程序时，Streams在运行时根据基于集群的负载均衡和可用性指标，自动决定在何处运行PE，让它能够重新配置运算符在其他服务器上运行，以便在服务器或软件出现故障的情况下确保流的连续性。如果希望对放置有更多的控制力，也可以通过编程方式指定哪些运算符在哪些服务器上运行，哪些运算符应该一起运行或单独运行，从而实现对一个或多个运算符的位置进行细粒度的控制。

这个自治流和可定制的平台让您可以轻松地增加对流执行分析的服务器数量，您只需要添加额外的服务器，并指定运算符在这些服务器上运行。Streams的基础架构确保，无论在不同的服务器上运行运算符，还是在同一个服务器上运行运算符，数据都可以成功地从一个运算符流到另一个运算符。您不仅可以添加或删除服务器，还可以动态地添加一些应用程序，它们会自动连接到正在运行的应用程序，也可以通过编程方式重新配置它们。当然，您也可以动态地删除应用程序，这样就可以更改优先级关系，并随着时间的推移改进分析。这些特性提供了高度的敏捷性和灵活性，让您可以从小规模开始，并按需扩展平台。

很像BigInsights，Streams对于结构化数据和非传统的半结构化或非结构化数据(来自传感器、语音、文本、视频和财务来源，以及许多其他高容量的来源)都非常适合。

由于Streams和BigInsights都是IBM大数据平台的一部分，您会发现可对运动数据和静止数据应用您构建的相同大数据分析会实现更高的效率。例如，通过Text Analytics Toolkit构建的提取器可以部署在Streams和BigInsights中。

InfoSphere Streams的工作方式

如前所述，Streams旨在对运动数据进行分析。在Streams中，数据以管道方式不断流过一系列运算符，就像巧克力工厂中流经生产流水线的配料。有些运算符会放弃没有用的或不相关的数据，就像可可豆分拣机可能会放弃太小的可可豆一样。其他一些运算符可能将数据转换成一个派生的数据流，就像将可可豆粉碎和液化一样。有些运算符合并不同类型的数据，就像以正确的比例将坚果混进巧克力混合物一样。

如果运算符的速度太慢，无法跟上数据流，可以拆分数据流并将其发送到这些运算符的并行实例，大致就像一个工厂可能安排其生产流水线连接到多台并行横塑机。有些运算符可能会发送不同类型的数据到不同的下游运算符，就像将带坚果的巧克力条发送到一个包装站，将不带坚果的巧克力条发送到另一个包装站。运算符甚至可能将信号发送到分析的早期阶段，以改变行为，非常像工厂中的质量控制，如果样品不符合规格要求，可能会增加坚果对巧克力的比例。

然而，只能在工厂临时停工的情况下才可以修改生产流水线，与之不同的是，不需要停止分析就可以改善、添加或删除Streams运算符。Streams对于高吞吐量和及时的分析是一个很好的方法；它让业务能够利用及时的智能实时地执行操作，最终为业务带来更好的结果。Streams提供各种运算符来将数据和结果存储在一个静止引擎中，发送操作信号，或者只是丢弃在动态分析中被认为没有价值的数据。

流是什么？

一个流是一个数据元素的连续序列。可将Streams应用程序看成是由有向边连接的节点图。图中的每个节点都是一个运算符或适配器，对来自一个流的数据进行处理。运算符可以有零个或多个输入，以及零个或多个输出。一个运算符的输出(或多个输出)连接到另一个运算符(或多个运算符)的输入(或多个输入)。图中将各个节点连接在一起的边代表数据流在运算符之间移动。运算符的每个输出都定义了一个新的流，其他运算符可以连接到该流。在管道中早期出现的运算符甚至可以连接到由“下游”运算符产生的流，让控制流能够在发现新洞察时修改上游运算符的计算。图6-1表示一个简单的流图，其中从一个文件读取数据，并将数据发送到一个称为functor的运算符(该运算符以某种可编程的方式转换所传入的数据)，然后将该数据馈送到另一个运算符。在该图中，流式传输的数据被馈送到一个split运算符，然后它将数据馈送到一个文件接收器或数据库(取决于split运算符内部发生了什么)。

在一个流中的数据元素被称为元组(tuple)。在关系型数据库的意义中，您可以认为一个元组类似于一个数据行。然而，当Streams处理半结构化和非结构化数据时，一个元组是一个代表数据包的抽象，这就是为什么我们将一个元组视为一个给定对象的一组属性。元组中的每个元素都包含该属性的值，它可以是一个字符串、一个数字、一个日期，甚至某种二进制对象，如一个视频帧。对于包含半结构化数据的应用程序，Streams应用程序通常从元组开始，每个元组包括少量元数据以及一个非结构化的负荷，而后续的运算符逐步从非结构化负荷中提取更多信息。

最简单的运算符每次处理一个元组。这些运算符可以根据元组的属性特征过滤元组，从元组中提取附加的信息并转换元组，然后将数据发送到一个输出流。由于流包含永不结束的元组序列，如何才能跨不同的数据流进行关联，对元组排序，或计算汇总？

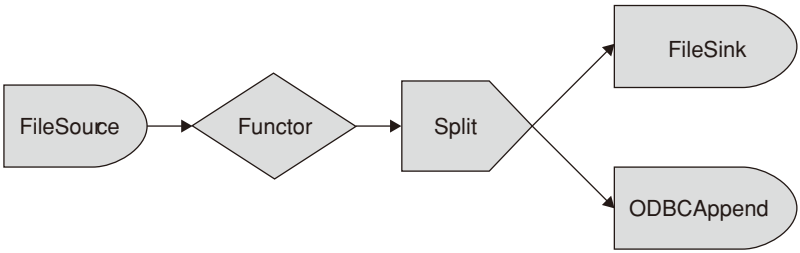


图6-1 一个简单的数据流，对数据应用转换，并根据预定义的逻辑将数据拆分成两个可能的输出

答案是数据窗口。数据窗口是元组的有限序列，看起来很像数据库视图。随着新数据不断到达，会始终更新窗口在，方法是消除最旧的元组并添加最新的元组。有很多种方式可以轻松地对窗口进行配置。例如，窗口的大小可以定义为N元组长或M秒长。窗口可以有多种改变方式，包括一次一个元组，或者立即更换整个窗口。每次更新窗口时，您都可以认为它是一个临时冻结的视图。可以很容易地将冻结视图与来自不同流的另一个数据窗口关联起来，也可以使用与关系型数据库中的汇总和连接类似的技术计算汇总。Streams中的窗口库为构建应用程序提供了令人难以置信的生产力。我们稍后会讨论窗口，并且会谈及各个运算符，但它是一个要理解的重要概念，因为Streams的目标不只是一次操纵一个元组，而是实时分析大型数据集并从跨多个元组、流和上下文数据的分析中获得洞察。

Streams还有一个复合运算符的概念。复合运算符包括一个可重用和可配置的Streams子图。从技术上讲，所有Streams应用程序都包含至少一个复合运算符(应用程序的主复合运算符)，但它们可包含一个以上的复合运算符(复合运算符可以嵌套)。一种复合运算符定义零个或多个输入流，以及零个或多个输出流。可将Streams传递给复合运算符的输入，并连接到内部子图的输入。内部子图的输出也可连接到复合运算符的输出。复合运算符可以暴露用于定制其行为的参数。嵌套复合运算符的一个极端示例是Streams中的Matryoshka示例应用程序，其灵感源于Matryoshka(或俄罗斯)套娃，即逐个缩小，一个套一个的著名木娃娃。

想像一个简单的示例，其中的一个复合运算符PetaOp包含一个子图，该子图由单个复合运算符TeraOp组成，然后，它又包含一个复合运算符GigaOp，以此类推。该示例如图6-2所示，它显示了只包含PetaOp运算符(在上面)与包含完全扩展的复合运算符(在下面)的应用程序部署对比。可以看到，复合运算符对于大型应用程序是非常强大和非常有用的，因为您可以封装和隐藏子任务，让开发人员能够专注于更大的目标。

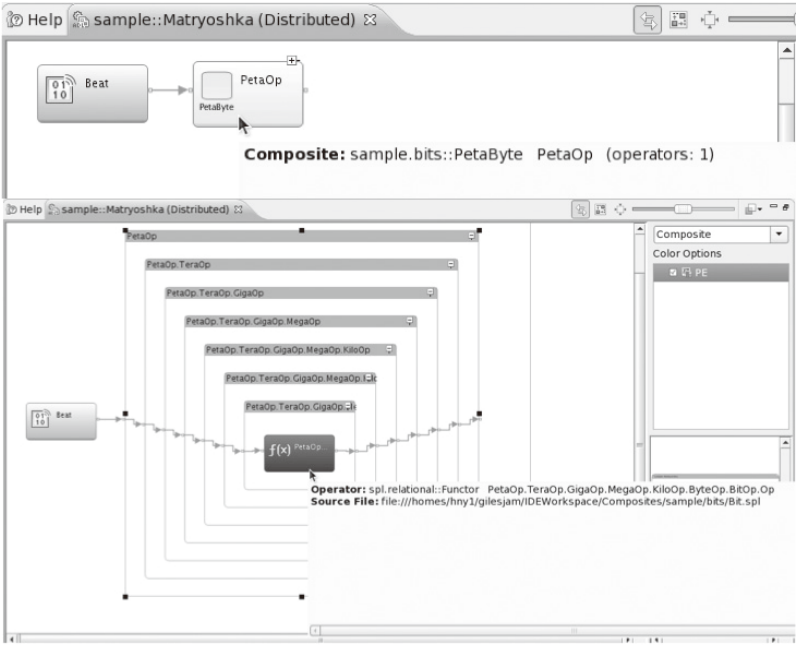


图6-2 Matryoshka示例应用程序包含上面的PetaOp复合运算符，下面是完全展开的PetaOp复合运算符(显示所有内部复合运算符)

Streams编程一点通

在过去的几年中，企业一直很满意Streams所提供的速度、可靠性和成本节省。此外，开发人员也已发现，分析工具包和编程工具在构建应用程序时提供了很高的灵活性。但是，使用Streams对于非程序员有点困难——就目前而言的确如此。

Streams 3.0完全侧重于易用性。Streams 3.0(以下简称为Streams)包括了一个可提供“按想像工作”的拖放式编程体验的集成开发环境(IDE)、新的运营和数据可视化、新的分析工具包、新的管理控制台特性，以及安装增强功能，这一切都有助于提供大幅改进的用户体验。并不是每家公司都可以一波接一波地聘请新程序员，这就是为什么大数据平台的可使用性及其功能是同样重要的。

Streams还让解决方案加速器成为最普遍的用例。例如，IBM提供一个名为IBM Accelerator for Telecommunications Event Data Analytics的可定制解决方案加速器，它使用Streams为电信处理呼叫详

细记录(CDR)。名为IBM Accelerator for Social Data的一个可定制解决方案加速器基于社交媒体为线索生成和品牌管理提供分析。您可以在第9章了解有关所有IBM大数据加速器的更多信息。

与Streams进行的大多数最终用户交互都通过Streams Console、Streams Studio或streamtool命令行界面来进行。我们将在本节的其余部分介绍这些强大工具的要点。

Streams Console

Streams Console是一个基于Web的工具，提供管理服务和有关Streams实例的大量信息。您可以快速查看Streams实例的健康和状态；管理和监控集群中的服务器、服务和应用程序；并控制实例的整体设置(如安全性)。此外，您可以在应用程序中配置运算符的Views，它们取样和缓冲数据，实现从Streams到可视化工具的实时导出。您还可以配置Charts，以便直接在Streams Console中可视化数据。

图6-3的底部显示了通过Streams Console实现主机(服务器)管理。一个可排序和可筛选的表显示了集群中每个可用的主机。您一眼就可以在Status字段中看到主机的状况。您还可以看到在每个主机上的服务状态、主机对于正在运行的应用程序是否可用，以及主机的度量指标，包括启用度量指标收集时的平均负载和CPU核心数量。可将IngestServer等标签添加到主机，帮助实现集群中的最佳应用程序放置。主机出现错误时，您只需单击用于确定问题的一个按钮就可以查看或下载日志。也可以让一台服务器上的工作负载保持静默，以便使其退出服务，进行维护工作。也提供作业、运算符和处理元素的类似特性；例如，图6-3的上半部显示了控制台的作业视图。

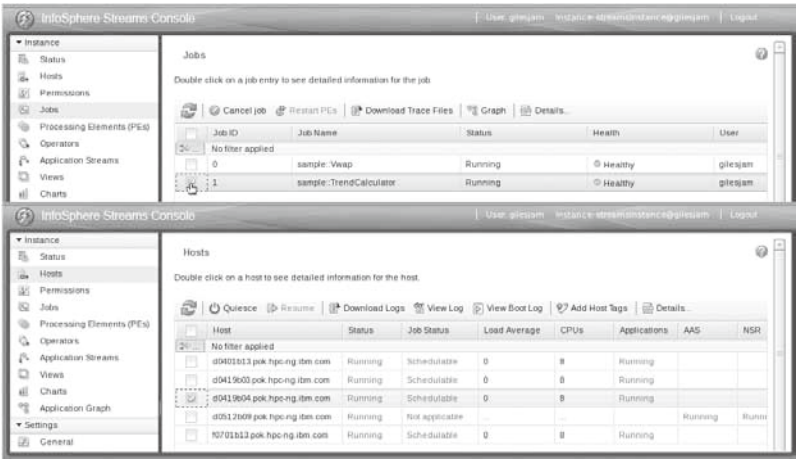


图6-3 Streams Console中的作业和主机管理

操作可视化

在Streams Console中有一个非常好的新特性，即可以使用图形以非常自然的方式监视Streams应用程序的结果。Application Graph特性在一个可定制的交互式窗口中显示所有正在运行的应用程序，以及来自Streams实例的运行时指标。

例如，图6-4展示了两个正在运行的金融服务应用程序，一个计算趋势指标，一个按每单交易计算每个股票代号的成交量加权平均价(VWAP)。在图6-4的上半部分，您可以看到该图形被配置为根据元组速率不断地更新运算符的颜色。还会更新线条的粗细，它与数据流动速率是成正比的，让您可以快速了解有多少数据流过。图6-4的底部显示了另一个视图，其中的运算符是按作业分组的。通过单击TrendCalculator应用程序中的一个运算符，可以显示额外的信息，如由运算符所处理的元组。单击其他对象(如端口或一个流)还可以提供大量详细信息。这些图形与以前Streams版本中的Streams Live Graph类似，但现在通过Streams Console也可以显示它们，让管理员可以更轻松地访问它们。

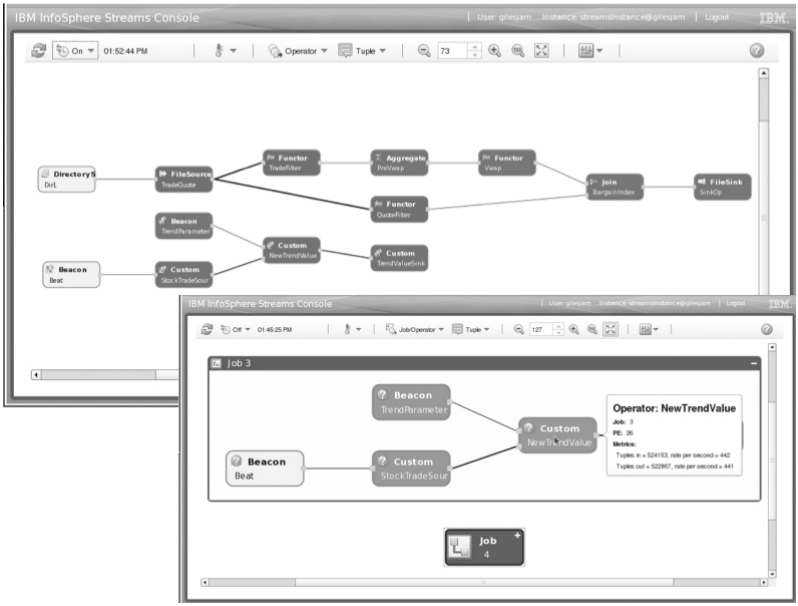


图6-4 Streams Console显示了一个Application Graph，其中有两个正在运行的作业，按作业分组，在下面显示了有关NewTrendValue定制运算符的详细信息

数据可视化

Streams Console其中一个最重要的新特性是数据可视化。除了让实时数据对外部可视化工具可用，还可以直接在Streams Console中对实时流式传输的数据进行可视化。通过选择和重新排列来自元组的属性，以及进行过滤，可以对表和图形进行广泛的配置。图6-5显示了IBM股票报价机的一个图表，其中包含VWAP、Min Price、Max Price和Average Price。(我们的律师希望我们提醒大家，这是来自一个示例应用程序的旧数据，所以如果您是一位短线交易者，请不要使用这些图表。)

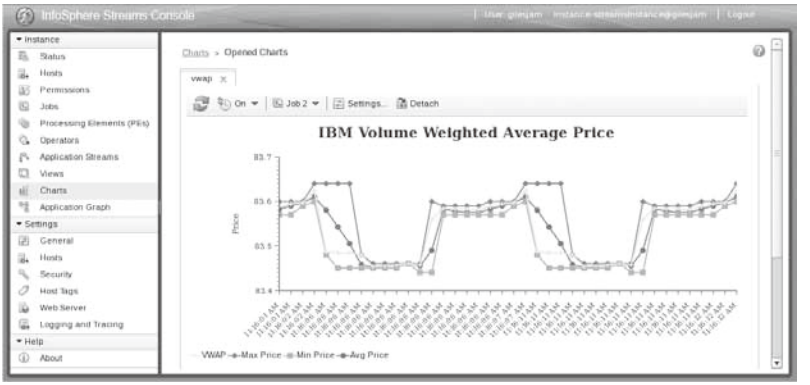


图6-5 图表显示代号IBM的VWAP计算

在Streams中可以找到的数据可视化小部件非常多；更多其他示例请参阅图6-6。

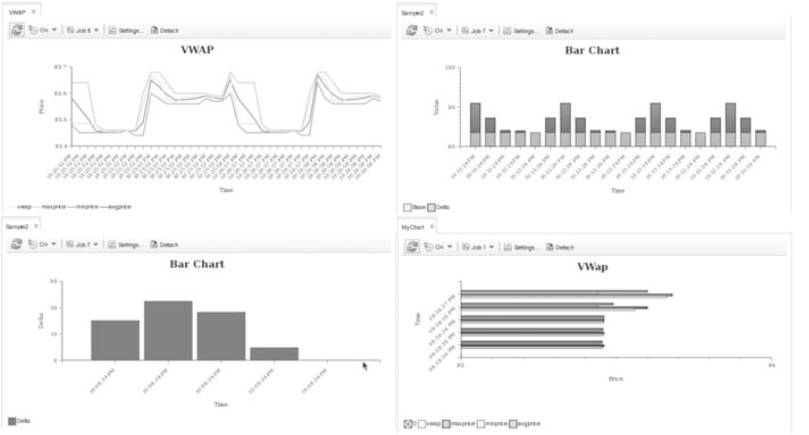


图6-6 在Streams Console中提供的图表示例

面向Streams的集成开发环境: Streams Studio

Streams Studio是一个交互式工具集，用于创建和部署Streams应用程序。这是一个典型的IDE，如果您有应用程序部署方面的经验，您会爱上这里所提供的特性: 内容辅助、设计时辅助、上下文相关的帮助、模板、设计时验证、拖放式调色板等。除了全新的拖放式应用程序构造和Streams Processing Language (SPL)编辑器，Streams Studio还可以配置和管理Streams实例。启动Streams Studio时，要注意的第一点是在First Steps下面的一个分步Task Launcher。构建Streams应用程序时，Task Launcher会引导您完成从概念设计到部署的一切步骤，以及这个过程中的所有任务。

Streams最明显的改进是全新的拖放式应用程序开发工具，称为Streams Graphical Editor。利用这个工具，无需编写一行代码就可以创建和部署Streams应用程序！认识到构建应用程序通常从在纸上画草图开始，Streams Studio允许用户在Streams Graphical Editor中绘制应用程序流程，而不必选择运算符实施。随后可以将通用运算符连接到一起并进行操作，直到流程正确无误。也可使用要执行的函数来注释运算符和图形。其他用户或开发人员可在以后选择和配置剩下的运算符实施，如果所有的工具包中都没有某些运算符，也可以创建新的运算符。

例如，图6-7显示了利用标记为Reader、Filter和Writer的运算符勾勒出的一个非常简单的应用程序。Reader的实施已知是一个FileSource(一个内置的运算符，用于从文件读取数据)。Filter和Writer运算符是通用运算符，它们用作占位符，直到选择了某种实施。在本例中，架构师已注释了图形，表示应使用Streams标准工具包Filter运算符并根据股票代号进行筛选，从而实施通用的Filter占位符。该图显示，一个用户搜索以“fil”开头的运算符，并将标准工具包Filter运算符拖动到图形上，以提供Filter占位符的实施。如果之后需要将实施修改为另一个运算符，也可以用相同的方式覆盖它。为Writer占位符选择一个运算符并配置属性后，一个完整的应用程序就可以运行了。

扩展现有的应用程序也易如反掌。图6-8显示了Streams自带的成交量加权平均价(VWAP)示例应用程序。在本例中，用户扩展了示例应用程序，将QuoteFilter流写入到一个文件。这是其中一个与Streams自带的所有示例有关的最大优势: 您可以扩展和自定义它们，从而构建您自己的应用程序。

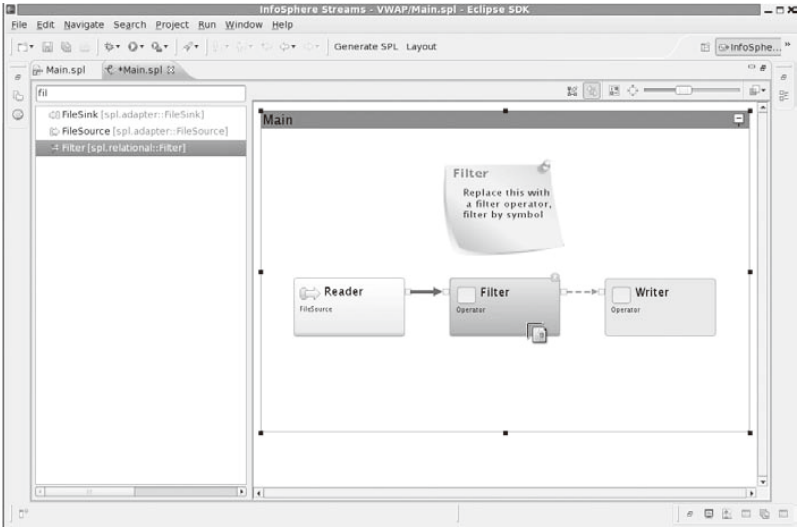


图6-7 使用Streams Graphical Editor，应用程序构建就是简单的拖放操作

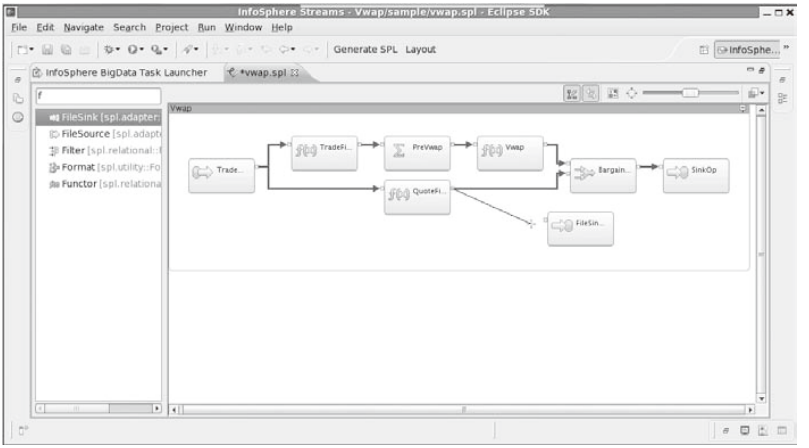


图6-8扩展VWAP示例，将来自QuoteFilter的元组写入一个文件

在本例中，用户只需将FileSink运算符从Streams标准工具包拖动到调色板，并将QuoteFilter流连接到新的FileSink运算符。FileSink运算符被重命名为QuoteWriter，并且file属性被设为QuoteFile.txt文件。运行新的应用程序之后，我们可以看到，在默认情况下，一次将一行元组属性写入QuoteFile.txt，其格式为一个逗号分隔值(CSV)格式，如下面的示例所示:

80.43,83.49,10,(1135711808,117000000,0),"IBM"

83.43,83.46,10,(1135711813,283000000,0),"IBM"

83.42,83.46,10,(1135711813,718000000,0),"IBM"

83.42,83.45,1,(1135711814,731000000,0),"IBM"

81.42,83.45,3,(1135711815,751000000,0),"IBM"

即使可以在Streams Graphical Editor中构建完整的应用程序，该工具集也提供了与Streams Processing Language编辑器的强大联动性，让高级开发人员可以进一步定制应用程序。例如，该集成让您能够实施最初并不存在于Streams工具包中的定制运算符。开发人员经常会发现，从Streams Graphical Editor开始，并且在勾勒最初的应用程序之后，在它与SPL编辑器之间来回切换是很方便的。图6-9显示的SPL编辑器带有我们使用Streams Graphical Editor添加的新QuoteWriter运算符。在清单底部高亮显示了新生成的代码。如果我们以后删除了那些代码，Streams Graphical Editor也会显示QuoteWriter运算符已被删除。

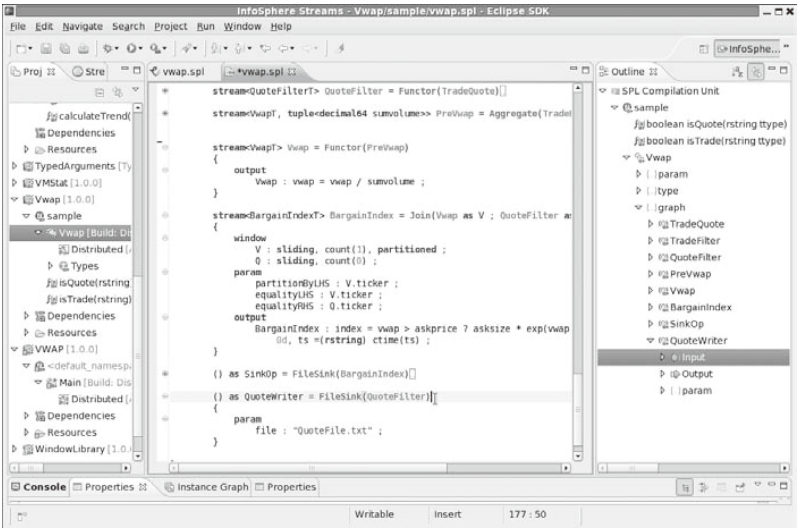


图6-9 Streams Graphical Editor和SPL编辑器是链接在一起的，可实现双向更新。

您还可以通过替换现有的运算符来更改应用程序。如果我们想在BigInsights中将QuoteFilter流发送到HDFS，而不是把它发送到一个正常的文件，我们只需从大数据HDFS工具包中选择内置的HDFSFileSink，并将它拖到FileSource运算符的顶部。一旦指定了新的实施，就可以配置HDFSFileSink运算符的独特属性，并且应用程序随时可以运行了。

Streams Studio可以部署应用程序，并在集群上运行它们，也可以在一台服务器或笔记本电脑上以单机模式运行它们。像Streams Console那样，Streams Studio也可以显示正在运行的应用程序的图形和度量指标。还有许多用于理解应用程序的其他特性，如根据作业、主机和流速率为图表添加颜色或进行分组的能力。

如果一个应用程序的行为与预期有异，也可以快速从集群收集日志并显示日志。例如，Streams Studio将颜色显示应用到一个作业及其运算符，甚至用颜色显示包含运算符的PE，不同的PE具有不同的颜色。

此外，通过单击任何运算符，都能够高亮显示上游或下游运算符，以了解出处(在流中应用到该运算符的数据和任何变更的起源)。如果运算符没有接收到预期的数据，能够迅速查看上游运算符对于调试大型应用程序非常有用。例如，如果运算符没有接收到数据，通过高亮显示上游运算符，您可以迅速回溯应用程序中的问题。图6-10显示了一个Instance Graph，其中正在运行VWAP和TrendCalculator作业。QuoteWriter运算符的上游运算符被高亮显示，以确定是什么生成了它的数据。

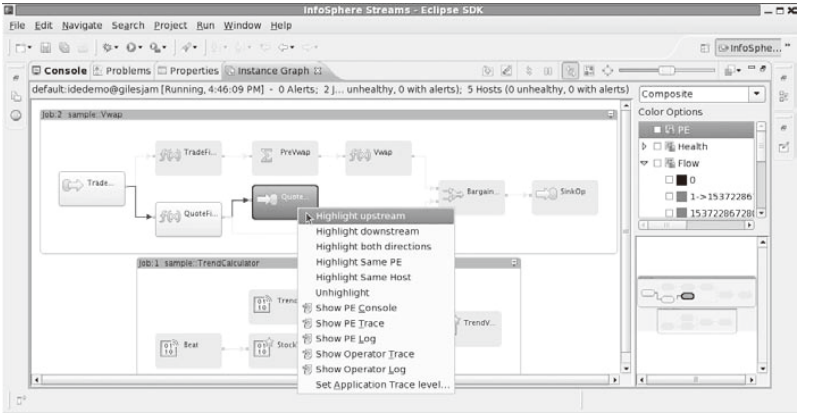


图6-10 Instance Graph中包含VWAP和TrendCalculator应用程序，QuoteWriter运算符的上游运算符被高亮显示。

Streams Studio中用于调试应用程序的最有用的新特性是，能够在Instance Graph中点击一个流来显示在该流上通过的数据。就像Streams Console中的Table视图，以可以过滤和操纵的表格形式显示在流上流动的实时数据。图6-11显示了一个用户为NewTrendValue流选择了Show Data。还显示了在用户运行完向导来配置应该应用的属性和过滤器后来自流的实时数据。可以将表中的数据暂停，以便在不影响正在运行的应用程序的情况下进行更详细的检查；在恢复更新后，会显示位于视图缓冲中的最新数据。因为当数据元素通过时，您可以很快地在流中看到它们，所以非常容易扩展和调试应用程序。

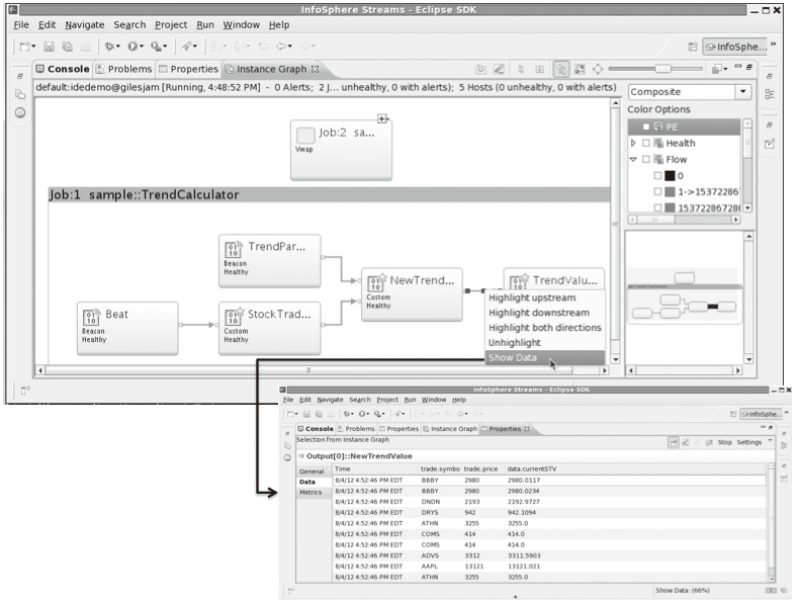


图6-11 通过直接单击Instance Graph，显示从NewTrendValue流馈送的实时数据

启动和运行体验: 安装和配置环境

集群的安装和配置可能极具挑战性。虽然我们不能详细介绍Streams安装程序和streamtool命令行工具，但我们已经通过大量投入来确保尽可能实现无缝的启动和运行体验。如果您曾经安装并配置过Apache的Hadoop集群，并将它与Cloudera或BigInsights的体验进行过比较，就会欣赏我们在这方面的工作。执行一个快速简便的安装程序后，用户可访问Streams First Steps配置工具。此工具极大地简化了各种复杂的任务，如设置ssh密钥、设置环境变量、配置公钥和私钥，以及验证安装。在许多情况下，用户只需单击任务，就会用最小的配置自动完成它。(您也可以使用streamtool launch –firststeps命令随时启动First Steps。)

如您所见，有许多功能强大的工具可以帮助用户和开发人员组装、部署、运行和调试Streams应用程序。还有工具可以有效地安装和管理集群，它们提供了详细的健康和指标信息。内置的运算符让组装应用程序很容易，无需编写程序。使用Streams Console和Streams Studio让可视化实时流式传输的数据变得比以往任何时候都要更容易。

流处理语言

使用新的Streams Graphical Editor，无需查看任何源代码就可以创建许多应用程序。然而，在每一个Streams应用程序下面的是流处理语言(Streams Processing Language，SPL)。SPL是一种结构化的应用程序开发语言，可用于构建Streams应用程序。Streams Graphical和Streams Processing Language编辑器均支持它，这些编辑器集成了“往返能力”，所以能够在两个编辑器之间来回切换。

SPL是一个非常强大的声明性语言，为构建Streams应用程序提供了比使用低级语言或API更高层次的抽象，与直接编写MapReduce应用程序相比，Jaql或Pig让Hadoop开发更容易，SPL也有类似的效果。Streams 3.0引入了一个全新的XML数据类型和运算符，以便在本机处理XML。源和接收适配器支持XML数据，使用新的运算符和函数可以对XML进行查询和操作，并且XML和Streams元组可以互相转换。此外，新的内置函数和运算符让SPL得到了广泛的加强；事实上，我们的合作伙伴之一，TerraEchos, Inc. 的创始人兼CTO Alex Philp博士指出，“由于流处理语言的敏捷性，他们的开发人员提供应用程序的速度可以快45%。”我们认为，瑞典排名第一的Uppsala University的教授Bó Thide在谈到SPL时说得最好，“Streams允许我再次成为一个空间物理学家，而不是计算机科学家。”毕竟，即使技术再好，但如果不能快速地将它应用到手头的业务需求，那又有什么意义呢?

对于利用Streams Graphical Editor构建或用SPL编写的Streams应用程序，可使用Streams编译器进行编译，该编译器将它们转换成二进制(bin)的可执行代码——此可执行代码在Streams环境中运行，以完成集群中多个服务器上的任务。SPL程序是我们在上一节所述图表的一个基于文本的表示：它定义源、接收器和运算符，以及它们与流互联的方式。例如，下面的SPL代码从一个文件中逐行地将数据读取到元组中，将每一行转换为大写，然后将每一行写入标准输出。

```
composite toUpper {
graph
stream<rstring line> LineStream =
FileSource() {
param    file          : "input_file" ;
         format        : line;
}
stream<LineStream> upperedTxt =
Functor(LineStream) {
output upperedTxt    : line = upper(line);
}
() as Sink = FileSink(upperedTxt)
{ param  file          : "/dev/stdout" ;
format                 : line;
}
}
```

在此SPL代码片断中，内置的FileSource运算符从指定的文件中读取数据，每次读取一行，并把它放进名为LineStream的流中，它只有一个元组属性，名为line。内置的Functor运算符使用LineStream流，将来自每一个流元组的line属性转换为大写文本，并使用与LineStream相同的元组格式创建一个名为upperedTxt的新输出流。然后，Sink运算符读取数据的upperedTxt流，并将元组发送到标准输出(STDOUT)。注意，应用程序被包装在一个复合运算符中，如前面的章节所述，它封装了函数；同样，所有应用程序都包括一个或多个复合运算符。

此代码片段代表了最简单的流，包含一个源、一个操作和一个接收器。当然，Streams的强大之处在于它可以跨大型服务器集群运行大规模并行作业，其中每个运算符或运算符组都可以在一台单独的服务器上运行。但是，在我们介绍Streams的企业级功能之前，先看看在此产品中提供的一些最受欢迎的适配器和运算符。如本章前面所述，Streams在标准工具包中提供了近30个运算符，在特殊工具包中提供了几十个运算符，如数据挖掘和文本分析，并为开发人员提供了数百个开箱即用的函数。

源和接收适配器

为了对数据流执行分析，数据必须进入一个流。当然，分析完成后，数据流必须去某个地方(即使那个“某处”被定义为无效，数据比特也被扔到“无处”)。让我们来看看可用于摄取数据的最基本的源适配器，以及可以将数据发送到哪些最基本的接收适配器。也

许我们所描述的最强大的适配器是import和export—这些运算符为可以在部署时和运行时进行配置的作业提供动态连接。

FileSource和FileSink

顾名思义，FileSource和FileSink是标准的内置适配器，用于从文件读取数据或将数据写入文件。使用参数来指定用于读或写操作的文件的名称和位置。另一个参数识别文件内容的格式，可能是以下格式之一:

- txt 简单的文本文件，其中每个元组都是文件中的一行
- csv 包含逗号分隔值的文件
- bin 包含二进制数据元组的文件
- line 包含文本数据行的文件
- block 由二进制数据块组成的输入流(很像一个BLOB)

还有许多其他可选参数，可用于指定列分隔符、行结束标记、分隔符、压缩等。

TCPSource/UDPSource和TCPSink/UDPSink

TCPSource和TCPSink适配器是在Streams中使用的基本TCP适配器，用于从套接字读取内容以及将内容写入套接字。使用这些适配器时，要指定IP地址(IPv4或IPv6)以及端口，适配器会读取套接字，并将元组生成到流中。这些适配器的参数与FileSource和FileSink适配器的参数在数据流格式(txt、csv等)方面是相同的。UDPSource和UDPSink适配器读取和写入UDP套接字，其方式与基于TCP的适配器相同。

Export和Import

export和import适配器配合工作，可连接Streams实例内的作业。export适配器可以让某个作业内的数据对已经部署或在未来可能部署的其他作业可用。您可以使用export适配器导出数据，并将一个streamID分配给导出的流，以及可选的名称/值对，以进一步描述流的特征。

为流分配这些导出属性之后，被部署到同一个Streams实例中的任何其他Streams应用程序都可以使用一个import运算符和一个与导出属性匹配的订阅表达式来导入此数据—假设应用程序有权访问已导出的数据流。使用内置的API，在运行时可以动态修改导出属性和导入订阅。这意味着，根据所提交的作业和所完成的处理，随着时间的推移可以用更好的方式构建应用程序。例如，利用参数textForSentimentAnalysis=true，情绪分析作业可以订阅所有流。可能有些作业处理电子邮件和博客文本会导出该参数。在未来的某个时间，如果您决定要处理情绪的即时消息文本，只需要部署作业来处理即时消息，并导出textForSentimentAnalysis=true参数—它将以高效的点对点方式自动连接到情绪分析作业。使用export和import是一个非常强大的方法，可

以在同一个Streams实例中所运行的应用程序之间动态地流式传输数据。

MetricsSink

MetricsSink适配器是一个非常有趣和有用的接收适配器，因为它让您可以建立一个命名仪表，每当一个元组到达接收器时，都会将其递增。您可以将这些仪表视为可以使用Streams Studio或其他工具来监控的量器。如果您曾经开车经过流量计数器(那些横跨在路口或道路上的黑色橡胶软管)，就会明白了。虽然流量计数器测量的是通过指定点的流量，但MetricsSink可用于监控流出数据流的数据的数量和速度。

运算符

运算符是Streams分析引擎的核心。它们从上游适配器或者其他运算符获得数据，操作这些数据，并创建要发送到下游运算符的新流和新元组(可能是直通的)。除了来自输入流的元组，运算符可以访问用于更改运算符行为的指标，例如在高负载期间。在本节中，我们讨论一些比较常见的Streams运算符，可以将它们串在一起，以构建一个Streams应用程序。

Filter

Filter运算符类似于实际的水流、熔炉或汽车中的过滤器: 其目的是只允许某些流式传输的内容通过。Streams filter运算符根据用户定义的条件从数据流中删除元组，该条件被指定为运算符的一个参数。以编程方式指定一个条件后，在运算符中定义的第一个输出端口将接收满足这一条件的任何元组。您还可以选择指定第二个输出端口来接收没有满足指定条件的任何元组。(如果您熟悉的提取、转换和加载[ETL]流，这类似于match和discard操作。)

Functor

Functor运算符从输入流读取元组，以灵活的方式转换元组，并将新的元组发送到输出流。转换工作可以操纵流中的任何元素。例如，您可以从一个流中提取出一个数据元素，并为通过特定functor运算符到达的每一个元组输出该元素的运行总计。

Punctor

Punctor运算符将标点符号添加到流中，然后，下游运算符可以使用它将流分离到多个窗口中。例如，假设一个流读取联系人目录清单并处理流过该流的数据。您可以使用punctor运算符，每当应用程序在流中观察到姓氏发生了改变，就将标点符号添加到流中，从而对联系人目录中的姓氏保持计数。然后，您可以在一个汇总functor运算符中使用此标点符号下游，以发出该姓氏的数量总计，之后将计数值重置为0，从而开始对下一组姓氏进行计数。

Sort

Sort运算符以指定的排列顺序输出它接收到

的元组。该运算符使用一个流窗口规范。仔细想想：如果一个流代表一个恒定的数据流，您如何对该数据进行排序？您不知道将要到达的下一个元组是否需要与作为输出被发送的第一个元组一起进行排序。为了解决这个问题，Streams让您可以指定一个操作窗口。您可以通过以下方式指定一个元组窗口：

- count 在窗口中包括的元组数量
- delta 等待，直到流中一个元素的给定属性被指定的增量修改
- time 允许进行窗口填充的时间长度，以秒为单位
- punctuation 用于分隔窗口的标点符号，由punctor运算符定义

除了指定窗口，您还必须指定一个表达式，定义您所希望的数据排序方式(例如，按流中一个给定的属性进行排序)。窗口填满后，sort运算符将根据您指定的元素对元组进行排序，然后以排列好的顺序将这些元组发送到输出端口。然后，窗口再次填满。默认情况下，Streams按升序排列，但您可以指定按降序排列。

Join

您可能已经猜到了，join运算符取出两个流，根据指定的条件匹配元组，然后将匹配的元组发送到输出流。当一个行到达一个输入流时，将匹配属性与第二输入流的操作窗口中已存在的元组进行比较，尝试找到匹配内容。正如在一个关系型数据库中，可以使用若干种类型的连接，包括inner joins(只传递匹配的数据)和outer joins(除了来自两个流的匹配元组之外，即使没有匹配也可以继续传递其中一个流的元组)。与sort运算符一样，您必须指定在每个流中要存储的一个元组窗口。

Aggregate

Aggregate运算符可用于总计窗口中各个元组的一个给定属性或属性组的值；该运算符也依赖于一个窗口选项将一组元组组合到一起。aggregate运算符让groupBy和partitionBy参数可以划分一个窗口中的元组，并对这些元组子集执行汇总。您可以使用aggregate运算符来执行count、sum、average、max、min、first、last、count distinct，以及其他形式的汇总操作。

Beacon

beacon是一个非常有用的运算符，因为它可用于动态创建元组。例如，你可以设置一个beacon，以按时间段(每n个十分之一秒发送一个元组)或迭代(发送出n个元组，然后停止)定义的不同时间间隔将元组发送到流中。beacon运算符对于Streams应用程序的测试和调试很有用。

Throttle和Delay

另外两个有用的运算符可以帮助您处理一个给定流的时间和流量: throttle和delay。throttle运算符帮助您设置数据流过一个流的“步伐”。例如，可将偶尔到达的元组以指定的速率(以每秒的元组数定义)发送到throttle运算符的输出。类似地，delay运算符可用于修改流的时间。可以设置一个delay运算符，在特定时间间隔之后输出元组；但是，如果使用delay，元组退出运算符的时间间隔与在元组到达时彼此间所存在的时间间隔相同。也就是说，如果元组A比元组B早到达10秒，元组B比元组C早到达3秒，delay运算符在元组被延迟指定的时间量之后，在退出时保持元组之间的这个时间间隔不变。

Split和Union

split运算符取出一个输入流，顾名思义，将该流拆分成多个输出流。该运算符读取元组中给定属性的值的一个参数化列表，并匹配元组的属性与这个列表，以确定将元组发送到哪个输出流。union运算符的行为正好相反: 它读取多个输入流，并将在这些输入流中发现的所有元组合并到一个输出流中。

Streams工具包

除了在前面的章节中所述的适配器和运算符，Streams还自带了一些工具包，支持更快地开发应用程序。这些工具包让您可以连接到特定的数据源，并操作在数据库或Hadoop中常见的数据；对时间序列数据执行信号处理；使用高级文本分析从文本提取信息；实时对数据挖掘模型评分；处理金融市场数据等。由于Streams工具包可以大大加快您使用Streams进行分析的速度，所以我们在这里会更详细地讨论Database、Big Data、Advanced Text Analytics和Data Mining工具包。该产品中还有更多工具包，如Timeseries、GeoSpatial、Financial Markets、Messaging、CEP、Internet、IBM PureData System高速适配器、DB2并行适配器和DataStage工具包。还有可以从developerWorks的Streams Exchange上免费下载的工具包(http://tinyurl.com/7sc6p8m)，如可用于图像处理的OpenCV工具包和HTTP连接工具包。

Database Toolkit: 面向关系型数据库的运算符

Database Toolkit让流可以从ODBC数据库读取数据或向其中写入数据，或从SolidDB数据库读取数据。此外，它提供了高性能的并行运算符，以写入DB2或IBM PureData System数据库。此工具包允许流查询外部数据库，以添加数据或验证流中的数据，然后进行进一步的分析。此Streams工具包中包括以下运算符:

- **ODBCAppend** 使用SQL INSERT语句将来自流的数据插入到一个表中

- ODBCSource 从表读取数据，并将每一行作为一个元组放进流中
- DB2PartitionedAppend 将数据插入一个特定的DB2分区，并且可以并行使用，以实现高性能加载
- NetezzaLoad 使用其高速加载器将数据插入IBM PureData System for Analytics (Netezza)
- SolidDBEnrich 从一个SolidDB表读取数据，并将该信息添加到流中的元组

Big Data Toolkit: 用于集成BigInsights的运算符

Big Data Toolkit拥有用于连接到Hadoop分布式文件系统(HDFS)的运算符。该工具包对于同Streams和BigInsights一起工作的应用程序是必不可少的。它支持高速并行写入到HDFS，实现最快的数据交换。此Streams工具包中包括下面列表中所示的运算符。

- HDFSFileSink 将来自流的数据写入HDFS
- HDFSFileSource 从HDFS读取数据并将它写入一个流
- HDFSSplit 将一个流拆分为多个流，让HDFSParallelWriter可以并行地将数据写入HDFS
- HDFSDirectoryScan 扫描一个HDFS目录，查找新文件，将文件名称写到一个流，供HDFSFileSource使用

Advanced Text Analytics Toolkit: 用于文本分析的运算符

Advanced Text Analytics Toolkit让您的应用程序可以充分利用在使用BigInsights时同样强大的文本分析函数。TextExtract运算符使用Annotated Query Language规范或分析运算符图形(AOG)文件，并处理作为元组到达的传入文本文档。然后它将结果作为元组发送给下游运算符。可以设置许多参数，如字典、语言和断词。该工具包对于社交媒体的实时分析很重要，并且是在第9章中所讨论IBM Accelerator for Social Data Analytics的一个重要组成部分。除了社交媒体，高级文本分析对于很多用例都非常重要，范围从日志分析(其中需要解析日志行，以提取其意义)到网络安全(其中的消息内容分析作为深度包检测的一部分)。

Data Mining Toolkit: 用于对数据挖掘模型进行评分的运算符

Data Mining Toolkit拥有对若干不同类型的数据挖掘模型进行实时评分的运算符。虽然数据挖掘软件(如SPSS)需要多次传递数据才可以构建模型，但评分往往可以在逐个记录的基础上完成。Data Mining Toolkit可以对按Predictive Model Markup Language (PMML)标准定义的数据挖掘模型进行评分。该工具包中包括下列运算符:

- Classification 支持对元组进行分类的

Decision Trees、Naïve Bayes和Logistic Regression算法

- **Clustering** 支持将元组分配给一个相关组的Demographic Clustering and Kohonen Clustering算法
- **Regression** 支持预测分析的Linear Regression、Polynomial Regression和Transform Regression算法
- **Associations** 支持关联规则，以预测因果关系

除了对PMML模型进行评分，Streams还可以对由SPSS Modeler Solution Publisher导出的大量算法进行评分。

企业级

过去构建的许多实时应用程序和并行处理环境来了又去。让Streams与众不同的是其企业级架构和运行时环境，它们足够强大和稳健，可以处理要求最苛刻的流工作负载；丰富的工具，可以构建、管理和部署Streams应用程序；数百个内置的分析运算符和函数；以及它与企业系统的集成。这是IBM及其研发部门带给大数据问题的价值。虽然有些企业拥有庞大的IT预算并尝试自己这样做，但将这些资金投入到核心竞争力和业务上岂不是更合理吗？

大型的大规模并行作业有着独特的可用性要求，因为在大型集群中，不可避免地会出现故障。好消息是，Streams拥有已考虑到这一点的内置可用性特征。再加上本书其他章节中所介绍的应用程序监控，Streams让您可以保持较低的管理成本和良好的业务声誉。与您的企业架构的其余部分相集成，对于构建一个全面的解决方案是非常重要的。这是在本书中反复出现的主题: IBM提供一个大数据平台，而不是一个大数据产品。在本节中，我们会介绍Streams可强化企业的哪些方面，包括可用性和集成。

高可用性

配置Streams平台时，需要指定哪些主机(服务器)将成为Streams实例的一部分。您可以为平台中的每个服务器指定三类主机，如下所示。

- **应用程序主机** 运行SPL作业。
- **管理主机** 运行可控制SPL作业流的管理服务(但不直接显式运行任何SPL作业)，管理实例内的安全性，监视所有正在运行的作业等。
- **混合主机** 可以同时运行SPL作业和管理任务。

在一个典型的较小环境中，将有一个管理主机，其余服务器用作应用程序主机，但对于大型部署，管理服务可以分布到多个主机上。

构建流式应用程序时，运算符被编译成一个或多个处理元素(PE)。PE可以包含一个或多个运算符，出于性能方面的考虑，在一

个PE内部往往将它们“融合”在一起。在Streams产品中，可将一个PE认为是物理部署的一个单位。来自同一个应用程序的PE可以在网络中的多台主机上运行，也可以在整个网络中与元组通信。如果有一个PE发生故障，Streams能自动检测到该故障，并从大量可能的补救行动中进行选择。例如，如果PE能够重新启动和重新定位，Streams运行时引擎会自动选择一个可用主机，并在其上运行PE；在该主机上启动PE；自动“重新连接”到其他PE的输入和输出(如果适用)。但是，如果PE仍然一遍又一遍地失败，并超出重试阈值(可能是由于经常性的基础硬件问题)，PE被置于stopped状态，并需要手动干预来解决问题。如果PE可以重新启动，但已被定义为不可重新定位(例如，PE是一个需要在特定主机上运行的接收器)，如果主机是可用的，Streams运行时引擎会自动尝试在同一主机上重新启动PE。同样，如果一个管理主机出现故障，假设您已经将系统配置为RecoveryMode=ON，则可以在其他地方重新启动管理功能。在这种情况下，恢复元数据具有在集群中的另一台服务器上重新启动管理任务所需的必要信息。

没有主机、网络或PE出现故障时，Streams中的数据处理是有保证的。您可能会想，如果这些组件之一出现故障，数据会发生什么变化？当一个PE(或它的网络或主机)失败时，如果没有特殊的预防措施，PE缓冲区中的数据(和PE被重新启动时会出现的数据)可能会丢失。由于最敏感的应用程序要求非常高的性能，在不同的硬件上部署两个或两个以上并行的流式应用程序实例是适当的——这样，如果某图形的一个实例中的PE失败，其他并行实例已主动处理所有数据，并且在已失败图形的恢复过程中可以继续。还可以使用其他战略，这取决于应用程序的需求。Streams往往被部署在生产环境中，高性能地处理每一位的数据，这是至关重要的，这些高可用性的战略和机制对于成功的Streams部署一直都是很关键的。事实上，Streams已针对企业部署进行强化，帮助一个客户从由停电引起的一次严重业务紧急情况中恢复。(如果一切都像Streams那么可靠就好了！)

集成是企业级分析的顶点

企业级解决方案的另一个方面是，它集成到现有企业架构的程度有多高。正如我们先前讨论的，大数据并非传统系统的替代品；它的存在是为了补充它们。协调传统流程和新时代的大数据流程需要供应商了解此等式的两端。Streams拥有丰富的高速连接功能，可以连接到各种企业资产，如关系型数据库、内存中数据库、应用服务器队列(如IBM WebSphere)等。虽然我们在第11章中会深入介绍在大数据世界中进行集成的详细信息和细微差异，但我们在本节先简要触及一些Streams特定的集成点。

Streams拥有多个接收适配器，支持将流

式传输的数据高速交付给BigInsights(通过BigInsights Toolkit for Streams)或直接交付给数据仓库进行静止数据分析。一个新的增强功能是能够利用IBM InfoSphere DataStage (DataStage)将Streams应用程序包括在ETL流中，并且让Streams应用程序能够将数据移进或移出DataStage。此集成让DataStage能够在流中嵌入分析，并且Streams能够访问通过DataStage提供的专用数据源和接收器。最后，Streams有能力把几乎所有来自SPSS Modeler Solution Publisher的分析流都融入一个Streams应用程序中，支持将丰富的分析工具纳入实时应用程序。

正如我们在整本书中所提到的，大数据问题需要分析静止数据和运动数据。Streams和BigInsights的集成为实时数据分析提供了一个平台(不只是产品)，并且为复杂工作负载提供了大量静止数据的分析。IBM为您提供了两全其美的解决方案，并且在结合这些产品时已考虑到安全性、企业服务水平协议的预期、产品和支持渠道在各国的本地化、企业绩效的预期等。

InfoSphere Streams的行业用例

为了让您简单了解一下Streams技术如何融入您的环境，在本节中，我们提供了一些行业用例，它们扩展了第2章中的示例。这些示例大多数是大数据平台示例，因为它们融合了大数据平台的其他组件与Streams。显然，我们不能在本书这么短的篇幅内涵盖所有行业，但我们认为本节会让您思考Streams技术为您的环境可提供哪些可能性，并为此感到兴奋。

电信

电信公司必须管理的呼叫详细记录(CDR)的数量是惊人的。此信息不仅有助于公司提供准确的客户计费，还可以从对CDR近乎实时的分析中收获丰富的信息。例如，CDR分析工作可以分析在其社交网络中“组领袖”的访问模式，从而帮助防止客户流失。这些组领袖所处的地位可能会影响其联系人从一个服务商转投另一个服务商。通过结合传统分析与社交媒体分析，Streams可以帮助您确定这些个人、他们所属的网络，以及对他们对谁能产生影响。

Streams也可用于推动实时分析处理(RTAP)营销活动管理解决方案，帮助提高营销活动的有效性，在更短的时间内将新的促销活动和软捆绑推向市场，帮助寻找新的收入来源，并丰富客户流失分析。例如，Globe Telecom利用从其手机收集的信息，针对每一个客户识别最佳的服务促销获得，以及提供该促销活动的最佳时机，这对其业务已产生了深刻的影响。Globe Telecom将新服务推出市场的时间从10个月缩短到40天，通过实时促销引擎大幅增加了销售量。

也可将适用于CDR的分析应用到Internet协议详细记录(IPDR)。IPDR提供了有关基于Internet协议(IP)的服务使用及运营支持可以使用的其他活动的信息，以确定网络质量，并检测可能需要维护的问题，避免它们导致网络设备出现崩溃。(当然，同样这个用例也可以适用于CDR)。涉及到CDR和IPDR处理时，究竟Streams的实时性和低延迟性如何呢？在一个需要不到10台服务器的示例中，我们支持的峰值吞吐速率相当于每秒500,000条详细记录，每天分析超过80亿条详细记录(是的，您看到的是正确的速率)，或每年超过5PB (5000TB)数据。除了所有这一切，还要对七天的数据(560亿条记录)执行传入记录的重复数据删除，并将延迟从12个小时缩短到几秒钟(是的，就是快了10,000倍以上)。Streams在网络监控方面保持每秒10GB的速率，在X射线衍射(XRD)方面保持每秒100MB的速率。如果这些示例对于您的环境而言都太小儿科了，值得一提的一件事情是，Streams具有网格可扩展性，所以您能够以弹性的方式动态地添加更多容量。诚然，Streams是一项改变游戏规则的技术。

执行、防御、监控和网络安全

Streams提供了改进执法和增加安全性的巨大机会，并对在此领域中可以构建的应用程序类型提供了无限的可能，如实时态势感知应用程序、多式联运监控、网络安全检测、合法监听、视频监控和人脸识别。企业也可以利用流分析来检测和防止流媒体网络日志和其他系统日志造成的网络攻击，以阻止入侵或检测在其网络中任何位置的恶意活动。

TerraEchos使用Streams提供隐秘的传感器监视系统，让拥有敏感设施的公司可以发现入侵者，避免让他们更接近建筑物或其他敏感设施。他们的技术已荣获多个奖项(其Fiber Optic Sensor System Boarder Application获得了Frost and Sullivan Award for Innovative Product of the Year等)。

金融服务部门

金融服务部门(FSS)及其子部门是一个很好的示例，说明了对流式传输的数据进行分析可以提供竞争优势(以及管理监督，具体取决于您的业务)。以极低的延迟，跨多个市场和国家同时对大量交易和市场数据进行分析的能力，为公司提供了微秒级的反应时间，通过套利交易和企业风险分析书带来了利润和亏损之间的差异。例如，在这一刻所发生的这一个交易对公司的风险状况有何影响？

FSS也使用Streams实现实时交易监控和欺诈检测。例如，Algo Trading在几个服务器上支持每秒约1270万条期权市场消息的平均吞吐率，并以50微秒的延迟为客户生成交易建议。

Streams通过无处不在的Financial Information eXchange (FIX)网关提供与一

个拥有丰富函数的库进行直接连接，帮助计算理论上的认沽和认购期权值。Streams甚至可以利用多种类型的输入。例如，在提供新服务时，FSS公司可以利用社交媒体来更好地了解客户。同样，信用卡公司和零售商也可以使用实时欺诈检测功能，以提供欺诈检测和多方欺诈检测(以及识别实时的向上销售或交叉销售机会)。

健康和生命科学

医疗设备旨在以极快的速度产生诊断数据。从心电图，到测量温度和血压的设备，到血液含氧量传感器等，医疗诊断设备会产生大量数据。驾驭这些数据并实时对其加以分析，可以提供不同于任何其他行业的好处。除了为公司提供竞争优势，在医疗保健行业中的Streams部署还可帮助您拯救生命。

例如，University of Ontario Institute of Technology (UOIT)正在多伦多建设一家智慧医院，它利用Streams提供一个新生儿重症监护病房，监视我们亲切地称为“数据婴儿”的新生儿状况。这些婴儿不断地生成数据: 每一次心跳、每一次呼吸、每一个异常，每秒都可能产生超过1,000条独特的诊断信息。Streams用作一个早期预警系统，帮助医生找出新方法，比以前提前多达24小时避免危及生命的感染。协同效应在这里也发挥了作用。可能某个单独监控的流符合正常的参数(血压、心率等)范围；但是，综合若干个流时，某个特定的值范围可能会成为即将发生的疾病的先兆。因为Streams对移动中的数据执行分析，而不仅仅是寻找超出范围的值，它不仅有可能挽救生命，也有助于降低医疗保健的成本。

我们无法在本书中介绍的其他用例……

我们不可能涵盖可受益于一个强大产品(如Streams)的所有用例和行业，所以我们在本节中将再简单介绍几个用例。

政府机构可以利用Streams广泛的实时分析功能来进行管理，例如通过监测和气象预报来管理野火风险，以及通过实时水流分析管理水质和水资源消耗。某些政府也正在其最拥挤的城市中利用来自出租车、交通摄像头和嵌入在道路中的流量传感器的GPS数据改善交通状况。这种实时分析可以帮助他们预测交通模式，并调整交通灯时间，以改善交通状况。

在公用事业行业中所产生的数据量正以爆炸式的速度增长。在整个现代能源网中的智能仪表和传感器以惊人的速度将实时信息发送回公用事业公司。

内置于Streams的大规模并行性让这些数据可以获得实时的分析，让能源生产商和配送商能够根据消费者不断变化的需求修改电网的容量。此外，公司还可以在分析流中包括关于自然系统(如天气或水管理数据)的数据，让能源交易商可以预测消费需求，并满

足客户的各种要求。这种方法可以提供竞争优势，并最大增加提高公司的利润。

制造商希望响应更灵敏、更准确并且具有丰富数据的优质记录，以及质量流程控制，从而更好地预测和避免既定的无法容忍的事件。电子科学领域(如气象预报、瞬态事件检测和同步原子研究)是Streams的其他机会。

从智慧电网，到文本分析，到“谁在和谁说话？”分析等，Streams用例几乎是无限的。

小结

在本章中，我们向您介绍了Streams背后的技术基础。我们谈到了Streams 3.0中使其比以往任何时候都更易用的所有全新增强功能。凭借一个更简单的启动和运行体验、引导任务完成的助手，以及通过丰富和敏捷的IDE实现的往返编辑，构建Streams应用程序比以往任何时候都要更容易。这是一个关键点，因为大数据技术的消费性对于大多数企业商铺而言代表了一条陡峭的学习曲线。考虑到在Version 2.0版中Streams Processing Language对Streams应用程序开发的影响，在Streams 3.0中的可用性增强应该是意料之中的。

Streams的关键价值主张是能够获得对业务前沿的分析——将典型的预测转换为一个临近预报。甚至很少有人在今天的大数据谈话中谈及这个价值主张，可以提供的信息甚至更少。Streams是经过验证的技术，其多个应用程序(参见本章中的详细介绍)可以帮助您的公司变成大数据时代中的一个领导者，因为您可以得到正确的答案，并比行业中的其他人更快地做出更好的决策。

释放大数据的力量

7 如果数据是新油田，那么您需要进行数据探索和发现

Vivisimo！和我们一起呐喊－使用意大利音调变化，同时挥动您的双手－就像您刚刚吃过一生中最好吃的正宗比萨饼一样！2012年4月，当IBM宣布收购Vivisimo时，我们满怀热情地这样呐喊。Vivisimo是一家致力于从多个数据源进行数据索引、搜索和导航的软件公司。

目前业务分析所面临的最大挑战之一就是组织将自己的数据存储在不同的信息孤岛上。因为这样做是有道理的：例如，我们将我们的交易数据保存在联机事务处理(OLTP)数据库中，将我们的电子邮件保存在Lotus Domino或Microsoft Exchange服务器中，并将我们的呼叫中心日志保存在SugarCRM等客户关系管理(CRM)存储库中。每个存储库都有特定的可用性要求、安全设置、服务水平协议(SLA)以及相关的应用程序。但是，如果要为特定的用户对来自您组织内各个数据源的所有相关数据建立一个完整视图，运气就不那么好了。有效支持您特定应用的基础架构而形成的信息孤岛会使得这种数据集成变得相当困难。毕竟，除非编辑图片，否则您不会将一张图片放大到250%然后检查一个点。放大图片并观看完整的图片有很大的价值，但这只能通过从许多数据源收集数据来实现。

IBM InfoSphere Data Explorer(原名为Vivisimo Velocity Platform－在本章的其余部分，我们将其简称为Data Explorer)代表IBM大数据平台的一个重要组件。Data Explorer技术让用户能够通过一个单一的集成视图访问他们所需的全部数据，无论数据采用什么格式、这些数据是如何进行管理的以及这些数据存储在哪里都无关紧要。能够在一个组织内从所有可用的存储库检索数据，是涉及大数据的分析操作的重要组成部分，对于探索性分析来说更是如此。(我们在第3章讨论了这一点：搜索和发现被列为IBM大数据入门的五种战略方法之一。)Data Explorer包括一个框架，让您可轻松开发业务应用程序，称为Application Builder。您可借助Application Builder建立可定制的基于Web的仪表板，为用户和上下文特定的界面提供Data Explorer能够抓取和索引的许多不同数据源。

Data Explorer让跨大数据资产的搜索工作变得更准确。底层索引较小(经过了压缩)，不

需要像其他解决方案那样经常进行维护，并且依据您的需求进行更细粒度的增量索引更新，不必更新所有数据。高效的索引文件，再加上动态扩展索引服务器的功能，让该解决方案成为高度可扩展的索引和搜索解决方案。Data Explorer还包括一个强大的安全框架，根据数据的原始内容管理系统中用户的安全配置，让用户只能查看他们有权查看的文件。

Data Explorer在最需要的时候提高了组织的生产率: 大数据时代的开端。该技术提供了一些技术来定位、保护并个性化业务数据的检索，已经帮助我们的许多客户实现了大数据的价值。

设想一下现在的大型喷气式客机，通常每架飞机都有多年来一直从事该行业的支持人员，就好像飞机是这些支持人员的孩子。现在回想一下您最后一次由于机械故障(本书作者多次遇到此类故障)而飞机里坐在门边很长一段时间的情景。航空公司呼叫该特定飞机的支持团队。航空公司可能做的最糟糕的是什么都不做，让飞机继续呆在那里——每分钟会造成成千上万美元的损失。

在这种情况下，从廊桥向客户支持团队拨打一个电话，无论是什么问题，支持团队都不得不努力解决。当然，这架飞机就如同一个客户——它有一个配置，有过去等。时间就是金钱，在这种情况下，机场费用、等候费、客户满意度及其他成本会随着时间的流逝而不断增加。与IBM合作的一个大型飞机制造商曾将信息锁定在独立的系统中，让支持团队几乎不可能访问所有的知识库，每个知识库都有一个不同的安全模式(数据被存储在SAP、FileNet、Content Manager、Siebel、共享文件里)。使用Data Explorer后，该大型飞机制造商建立了一个单一访问点来访问他们所有的存储库，同时对数据进行无缝、更细粒度的安全控制。多个前端应用程序使用一个共同的后端基础架构来检索数据。支持团队由于能够放大和缩小眼前问题的范围，因此提高了工作效率，从而减少了为一架飞机提供支持工作所需的人员。这能够让客户实现更好的收益，因为可以使用现有团队为新飞机提供支持，而不必在交付新飞机时雇用新员工。最后，借助Data Explorer，他们还让咨询台解决问题的延迟时间减少了70%，从而节省了数百万美元的成本并提高了客户满意度。

使用InfoSphere Data Explorer从多个源索引数据

Data Explorer是一个搜索平台，可以从多个源索引数据，并提供了单一的搜索界面，让用户能够在组织内外查看所有相关数据。尽管Data Explorer处理索引和搜索工作，但数据本身仍保留在原始数据源。(这个将功能带到数据处的模式也是Hadoop背后的原则之一。)图7-1展示了Data Explorer的主要组件的架构布局。

连接器框架

在制定搜索策略时，首先确定需要访问的数据源。Data Explorer包含了一个连接器框架(Connector Framework)，该连接器框架支持30多种常见的数据源，包括内容管理存储库、CRM系统、wiki、电子邮件存档、供应链管理存储等。

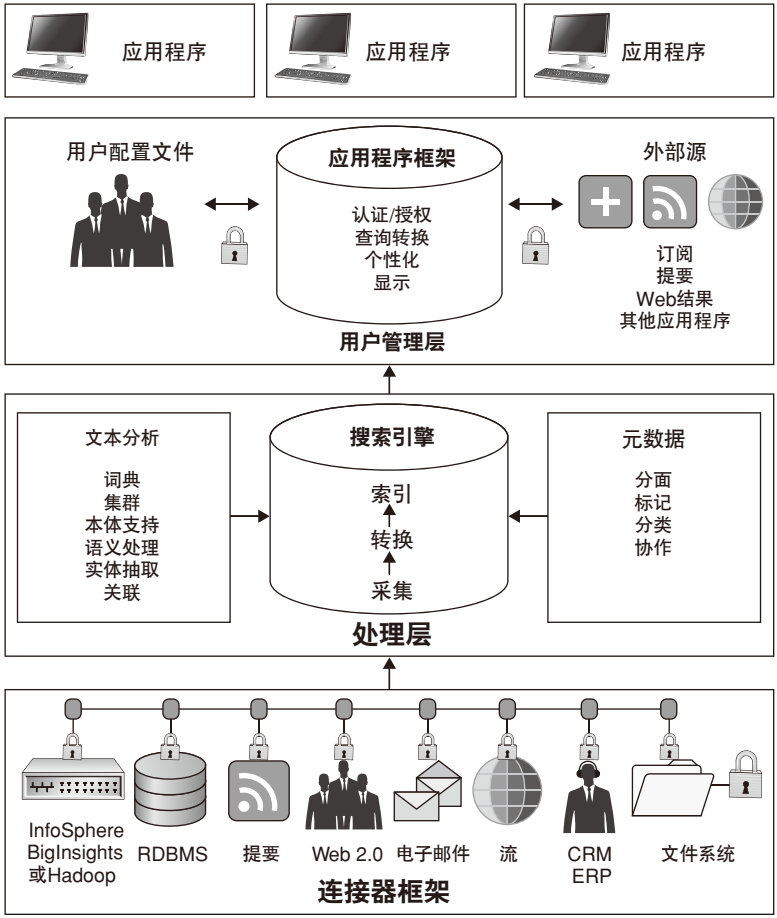


图7-1 Data Explorer架构

也有连接到InfoSphere Streams和InfoSphere BigInsights的连接器，显示了组件在IBM大数据平台中的深度集成。如果您的数据源没有连接器，请不要担心；Data Explorer还包括一个成熟的框架来建立可连接到专用数据源的其他连接器。

连接器框架利用支持的数据源来处理数据，以便进行索引。我们希望清楚地说明Data Explorer不在数据源管理信息；它只保存可用内容的索引，用于搜索、导航和可视化工作。

在许多情况下，人们都依赖于数据储存，如组织内外的Web存储库。您可以使用Data Explorer将这些远程源添加到已经为您的内部源建立的统一搜索环境中。Data Explorer不索引这些远程站点，但与远程站点的搜索引擎对接，将查询传递给远程站点的搜索引擎。然后接收结果集，对其进行解析，并将它们和来自本地数据源的数据一起呈现给用户。

一个复杂的安全模型让Data Explorer能够将源系统中用户对数据的访问权限控制映射到搜索引擎中对索引文件的权限控制，并在用户访问数据时行使这些权限。这个安全模型扩展到各个文档的字段级，这样，可通过一个文档内的段落或字段级别的权限控制来更好的保护文档的安全性，同时可以进行增量的更细，而无需为整个文档重建索引。因此，用户直接登录到目标存储库后，只能看到对他们可见的数据。例如，如果一个内容管理系统的字段级安全监管对预估盈利报告的访问，那么它可能允许一个特定用户访问执行摘要部分，但不允许该用户访问税前收入(PTI)等财务细节。很简单，如果不使用Data Explorer原则上看不到数据的话，那么借助Data Explorer也无法看到数据。

Data Explorer连接器检测何时添加或更改目标数据源中的数据。通过这些连接器，连接器框架能够确保索引文件反映目标系统中信息的最新视图。

Data Explorer处理层

Data Explorer处理层有两个目的，其中每个都反映了不同的阶段: 在内容可用时对其进行索引，以及处理来自用户和应用程序的搜索查询。在本工作流程的开始，连接器框架让来自各个存储库的数据可被爬取。在解析完数据后，使用一些不同的分析工具来转换和处理数据，包括实体提取、标记和用于分面导航的元数据提取。在整个数据运算阶段，处理层都保留来自所连接数据源的内容的索引。如果您的企业拥有描述您的数据集的现有信息，如分类法、本体论及其他知识表示标准，那么也能将该信息包含到Data Explorer建立的索引中。

从目标数据源收到的安全信息被处理层采集，并包含在Data Explorer为各个目标数据源建立的索引中。这样会启用我们前面所述的基于细粒度角色的安全功能，确保根据用

户对各个目标数据源的安全权限，让用户只访问他们有权查看的信息。

与IBM大数据平台的其他主要组件一样，Data Explorer的设计宗旨是基于针对海量数据的多节点并行处理架构,提供良好的扩展性。该产品已在生产环境中被用来索引上万亿的记录和PB级的数据。

从高可用性的角度来看，Data Explorer服务器具有主-主复制和故障切换功能。一台服务器脱机时，所有搜索和采集流都被重定向到其他的可用节点上。原始服务器重新联机后，其各个集合都自动与对端同步。如果某个集合已被损坏，则会自动恢复它。对于计划内的停机，Data Explorer服务器可在不中断服务(索引或搜索)的情况下进行升级、更换或取消配置。

秘密武器: 位置索引

索引是任何搜索系统的核心，也是影响查询性能的一个主要因素。在大数据实现中，索引结构、规模、管理和其他特点方面的差别会被放大，因为数据的规模在不断增加，复杂性也在不断提高。Data Explorer有一个明显的优势，因为它具有独特的位置索引结构，这个结构比目前市场上的其他搜索解决方案具有更全面的功能和更优异的性能。

要真正理解为什么位置索引让Data Explorer成为卓越的企业搜索平台，需要了解常规索引(被称为向量索引)的局限性(参见图7-2)。

使用向量索引方式时，会根据所有被提取的词汇在文档内出现的频度对其进行权重衡量(权重与频度成正比)。在查询时，这个权重衡量也受到搜索词汇与整套文档有关的独特性的影响(权重与在整套文档中出现的次数成反比——简言之，如果一个词不经常发生，那么这个词就是“特殊的”。)频度和独特性之间的平衡非常重要，像the这种经常使用的词可能在独立文档中经常出现(在整套文档中也经常出现)，要让这样的词不会影响搜索结果。

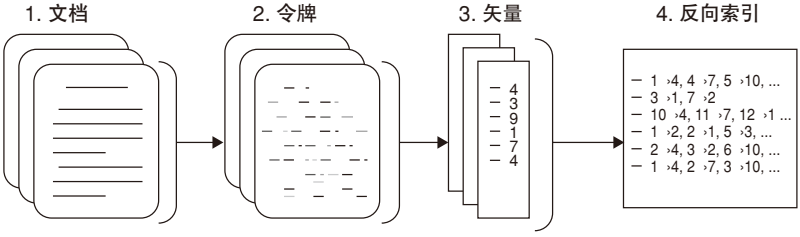


图7-2 向量索引

使用整套文档的结果来建立倒排索引，倒排索引把各个词及其权重映射到它们在文档中的位置。如果使用向量索引向搜索引擎发出查询，则会使用搜索查询中的词汇来计算类似的矢量。矢量与搜索词汇的矢量匹配度最高的文档就会被包含在排名第一的搜索结果中。

矢量空间方法有一些局限性，大多数局限性都是因为在文档被缩小为矢量后，不可能重建全部文档流。例如，不可能将文档的一部分看作独立的单元，所提供的关于此类文档的惟一线索是它被索引词汇的出现词频和独特性。

大数据和现代搜索应用程序需要的不仅仅是关于词频和独特性的信息。需要定位信息才能高效地执行短语或近似(proximity)搜索、使用近似作为排名因素或生成动态摘要。因此，在跟踪文档位置时(例如，多个搜索词汇在一个文档内的近似性)，传统的索引近似方案需要在矢量空间索引之外再创建一个结构，通常是文档特定的位置空间索引。当然，像生活中的大多数事情一样，没有什么是免费的: 这个额外索引的代价是需要更多的时间来索引文档，所产生的索引需要较大的空间量。

如前所述，Data Explorer也使用位置空间索引，但此处的区别是没有基本的矢量空间索引。位置空间索引比传统的矢量空间索引更加紧凑，因为Data Explorer只使用一个高效的结构，而非使用两个效率较低的基于文档的结构。 在位置空间索引中(参见图7-3)，文档被表示为一组令牌，每个令牌都有一个开始和结束位置。令牌可以是一个词或内容范围(例如，标题或作者的名字)。用户提交查询后，搜索词汇会匹配一段令牌，而不是整个文档。Data Explorer不计算矢量表示，而是将所有位置信息都直接保存在其索引中。这种表示法可完整地重建源文件，以及任何子部分的操作。

在大数据部署中，由于被索引的数据量较大，因此索引大小是一个大问题。许多搜索平台，特别是有矢量空间索引方案的搜索平台，所产生的索引是原始数据大小的1.5倍。Data Explorer高效的位置索引结构可产生紧凑的索引，并且会压缩该索引，让索引大小成为业界中最小的。此外，与矢量空间索引不同，位置空间索引不会在数据发生变化时增加；只有添加新数据时，位置空间索引的大小才会增加。

位置空间索引的另一个优势是字段级更新，在字段级更新中，修改文档中的一个字段或记录只会修改要重新索引的文本。如果使用矢量空间索引，则需要重新索引整个文档。这样会将频繁更新的系统中过多的索引负载消除掉，并近乎实时地将少量更新提供给用户和应用程序，变更虽少，但通常非常重要。

字段级安全性的概念与字段级更新有关，对

于智能应用程序非常有用，因为它让一个单一的分类文档能够包含不同级别的分类。

Data Explorer可将安全性应用到文档内的文本分段，包括字段。

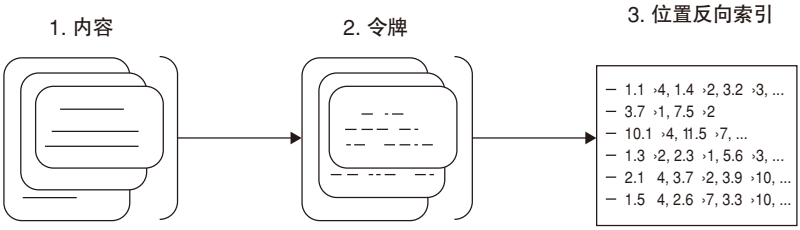


图7-3 位置空间索引

一个给定的文档虽然只被索引了一次，但根据用户的安全设置，可能对不同的用户群看起来不一样(回想一下本章前面的业务计划示例)。如果使用矢量空间索引，则无法进行这样的安全设置，因为用户要么访问整个文档，要么完全不能访问，这会限制组织在整个企业内共享信息的灵活性。

索引审计

对于需要进行详细审计和跟踪的大数据应用程序来说，被收集到索引中的数据可通过Data Explorer的审计日志功能进行完整地审计。Data Explorer为各条发送后进行索引的内容生成一个审计日志条目。保证日志条目包含在抓取和索引过程中遇到的所有错误和警告。借助审计日志，Data Explorer部署可确保总是能考虑全部内容: 各个项目要么被索引，要么已经触发了一个错误，然后将错误报告给管理员。这种内容完整性审计对于许多法律和合规性发现应用程序来说都是一个关键的要求。

用户管理层

用户管理层包括用户和应用程序与被Data Explorer索引的数据进行交互所需的全部资源。最重要的是，还包含一个连接到Data Explorer搜索引擎的接口，处理所有来自已连接源的数据请求。查询任何数据之前，用户必须进行身份验证。用户的身份以及各个源系统所建立的索引文件访问权限对应关系都存储在用户配置文件中，或登录时通过目录服务访问该信息(我们要说明一下，如果您使用LDAP或Active Directory服务，会自动获取此信息。)用户登录后，可以获得一个反映其配置的个性化界面。

该层还包含Data Explorer将查询联合到外部源的功能，这些外部源本身不被Data

Explorer进行本机索引，如互联网上付费订阅的信息服务。可将结果与本地Data Explorer结果合并，以创建相关信息丰富且扩展的视图。

Data Explorer让最终用户能够评论、标记和评价内容，以及为他们想与其他用户共享的内容创建共享文件夹。这个用户的所有反馈和社交内容都被馈送到Data Explorer的相关性分析，确保将最有价值的内容呈现给用户。用户还可以对他们的搜索结果发表评论。这些评论受到字段安全功能的保护，只有用户拥有适当的权限时才能创建或查看评论。此外，拥有相应权限的用户可以将结果保存到文件夹，这些文件夹可以是个人的，也可以在群组级共享，或在整个企业内共享。这就打造了一个强大的协作环境，可根据用户的活动在搜索结果中返回用户的配置。假设一个名叫Anna的用户向各种文档添加了评论和标记，包括Hadoop一词。那么任何包含Hadoop一词的搜索查询都将返回Anna的用户配置，即使她的配置数据和工作描述中没有提及Hadoop也是如此。

加强InfoSphere BigInsights

由于Data Explorer现在是IBM大数据平台的一个集成组件，因此其企业级索引和搜索功能也适用于InfoSphere BigInsights (BigInsights)。尽管BigInsights采用的Hadoop技术拥有强大的功能，可对大量结构化和非结构化数据运行复杂的工作负载，但对于有些用例，Hadoop并非可行的解决方案。Hadoop分布式文件系统(HDFS)面向大规模的批处理操作，可对数据集当中的大部分信息或所有信息展开工作。然而，在HDFS中涉及小数据子集的查询则表现不佳。通过检索已存储在HDFS中的内容，Data Explorer提供了一种方法来满足快速响应需求，同时不会影响BigInsights的强大功能。

此外，提取、识别和充分利用元数据的功能可大大提高搜索精度、可用性和相关性。事实上，通过识别自然语言文本中的实体和其他重要词汇，搜索过程可将结构添加到非结构化内容中。Data Explorer还添加了语义功能，如分类、聚类和分面导航等，让您能够按主题浏览搜索结果，或在不输入查询的情况下导航到单个结果。

带有视图的应用程序: 使用InfoSphere Data Explorer Application Builder创建信息仪表板

在任何大中型企业中，都有多个存储业务数据的存储库，员工总是要花费大量时间寻找信息。对于许多面向客户的角色来说(如销售或产品支持)更是如此，人们对许多系统的依赖性非常强，这些系统包含重要的商业信息，如CRM数据、产品知识和市场数据。

Data Explorer包含一个极有吸引力的工具: Application Builder应用程序框架。Application Builder让您能够为员工需要

访问的许多相关数据源建立一个多功能仪表板。虽然Data Explorer提供了原始功能来索引许多不同的数据源并高效地处理查询，但您可以使用Application Builder创建前端仪表板，这些仪表板可以是所有信息的展示界面。图7-4展示了Application Builder和Data Explorer协同工作的架构。Data Explorer的所有强大功能，如高性能搜索和保存用户特定的数据访问，都可在Application Builder加以利用。

Application Builder应用程序可用的选项是无限的。例如，您可以联合基于互联网的数据源来引入新信息，或收集来自证券市场的财务数据或Twitter等社交媒体数据源的数据。Data Explorer的社交功能还体现在以下方面：用户可通过集体数据标记工作来实现协作，并通过意见和建议共享知识。这些意见和建议适用于存储在Application Builder汇总的数据源中的数据。

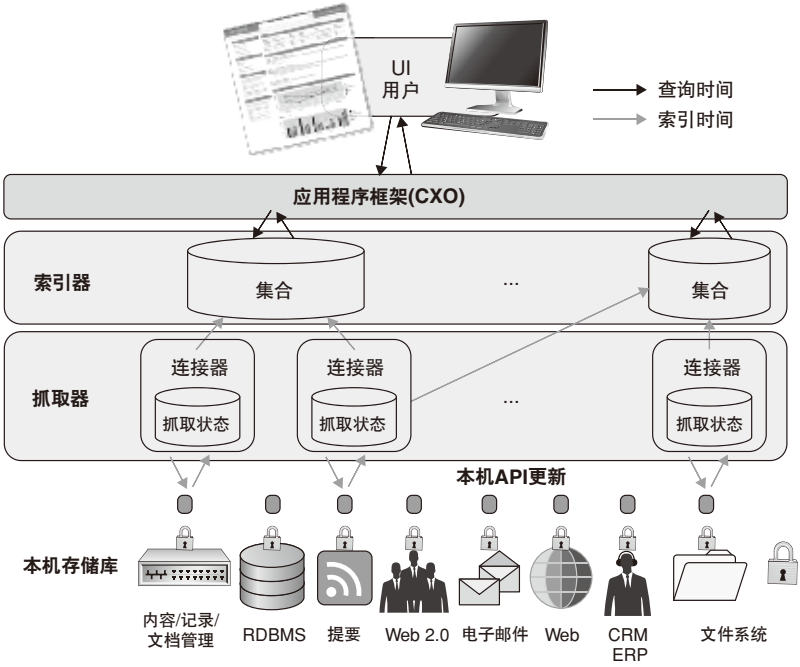


图7-4 Application Builder架构

Application Builder与BigInsights和InfoSphere Streams (Streams)进行了集成。BigInsights和Streams都能将数据传输到Data Explorer，然后Data Explorer将这些数据联合提供给Application Builder用户。此外，Application Builder也可以直接使用BigInsights或Streams的数据。例如，Streams可作为Application Builder中的实时数据源。除了托管为用户定制的内容外，Application Builder还提供了Data Explorer搜索界面。例如，分面搜索、聚类搜索和推荐引擎都可以在Application Builder中使用。

用Application Builder创建的仪表板不仅仅从不同的数据源收集数据。真正的功能来自于用户与Data Explorer中可用的数据集之间所定义的实体关系连接。有了这种连接，您的仪表板可提供来自所有这些数据源的、与您的用户相关的信息，而无需搜索这些信息。例如，您的支持人员可以定制仪表板，显示与他们自己的客户帐户有关的数据，如客户的购买历史记录、公开支持票和订购状态。这还包括来自于外部数据源的数据，如与客户或他们的股票价格有关的新提要。

总之，Application Builder让您能够定制员工所需信息的集成视图，搜索工具提供来自多个数据源的数据导航。这种结构的优势在于用户现在不仅能够在他们的企业内统一访问信息，而且在许多情况下，他们甚至不需要搜索这些信息，因为这种技术“把各个点连接起来”，并自动为用户提供相关的信息。

小结:

Data Explorer释放了大数据的力量

Data Explorer帮助组织充分发挥并优化其所有信息的商业价值，无论应用程序或数据源是什么都可以。它提供业界领先的大数据索引和搜索技术，并包含一个功能丰富的界面，让您能够轻松构建并向最终用户群推出个性化仪表板。Data Explorer是针对大数据时代从内向外进行设计的，采用了位置索引技术、细粒度权限控制等功能。有了Data Explorer的连接框架，数据可在被管理时保存在信息孤岛，同时数据专家、研究人员和商业用户可以专注于查询关键的问题。

在大数据世界，需要强大、准确、敏捷、灵活地搜索所有数据集，这比以往任何时候都更重要，因为数据的类型和量级发生了变化，数据量比以往任何时候都要大，也以更快的速度到达您的门口。根据我们的经验，许多大型组织都要为不知道它们应该了解的事情负责任，它们拥有大量的数据集，但与用户的连接却不全面，导致在进行搜索时混乱无序，只能去碰运气。Data Explorer是一个转折点——对于需要释放被困在各种信息孤岛中各种信息价值的组织来说，这是一个好消息。您也会这样呐喊的！Vivisimo！

大数据分析加速器

8 利用文本分析让自己与众不同

虽然大数据分类往往可分为结构化、半结构化和非结构化，但我们想提出的概念是，所有数据都有某种形式的结构(使用智能手机拍摄照片就可能附加了位置感知、时间戳、其格式和大小元数据等等标记。引用这些不同程度的结构代表着对数据进行分析和解读的相对容易程度；通常结构程度越低，就需要越多工作来提取洞察。例如，Facebook帖子是结构化数据，它被包装为Java Script Object Notation (JSON)格式。然而，它是结构的表示法内的自由格式文本，是非结构化的部分，并且是数据集内最难分析的部分。我们已经相当擅长于分析数据库中的信息，但那些数据已被清洗和蒸馏为高度结构化的形式。企业如今面临巨大挑战的地方是，要分析尚未被很好地格式化的数据，如电子邮件、法律文档、社交媒体消息和日志文件。由于组织越来越多地依赖于锁定在各种形式的文本数据中的信息，关键是向他们提供一个框架，不仅帮助他们了解在此文本中的内容，也可以帮助他们以具有成本效益的(即非高度专业化的技能集)和相对快速的方式来实现这一点。

有很多问题领域的特征都是非结构化和半结构化数据。我们认为文本分析可以是改变游戏规则的因素的其中一个领域是欺诈检测。要了解有关证券交易，保险索赔，或抵押申请等事务的完整故事，需要对围绕这些事务的所有数据进行分析。除了关系数据之外，这总是包括原始文本(电子邮件或表单中的文本字段)和半结构化数据(来自日志文件)。

数据篡辑是文本分析平台大幅提高学科能力的另一个领域。如今所存储的大量数据包括个人身份信息(PII)，这在许多国家中都必须得到充分的治理，以保护公民的隐私。为了确保合规性，该数据(例如，来自表单、记录或法律文档)必须被篡辑，以隐藏关键要素，否则将会泄露人员的身份。事实上，IBM InfoSphere Guardium Data Redaction产品针对此明确目的采用了作为IBM大数据平台核心的文本分析技术。

对于以“企业对消费者”为导向的公司，特别是在服务行业，能够全面了解每个客户的帐户是很重要的，这是客户关系管理(CRM)分析的范畴。许多有价值的活动(如，有针对性的市场营销和客户流失预测)都依赖于对客户行为的了解，因为这不仅涉及其事务历史，还涉及其呼叫中心交互的转录，甚至涉及其访问企业Web存在的点击流日志，所以需要文本分析来找到“下一个档位”。

在社交媒体分析领域，文本分析可以促进大数据项目的方式很明显。随着Facebook等在线社区和Twitter等微博服务的出现，人们以我们从来没有见过的大规模来公开地表达其个人感觉。企业可以更好地了解每个客户，不仅了解他们在社交媒体中说什么，还了解为什么他们都这么说。此外，知道各组别的人在社交媒体上所说的内容，这可以彻底改变营销人员如何评估其营销活动的覆盖范围和响应。

很明显，人们在听到短语“文本分析”时很快就会想像到情绪用例，但我们希望非常确定，您明白它还有更加丰富的内涵；包括盗版检测、情绪分析、投资研究，等等，大数据平台需要有完善的文本分析生态系统。

什么是文本分析?

对于我们刚才提到的每一个示例场景，其挑战是分析文本，找到正在搜索的元素，理解其含义，并以结构化形式提取它们，以便在其他应用程序中使用。IBM在该领域中有丰富的经验，我们已经亲眼见过许多组织尝试自己开始。因此，我们可以告诉您这并不是一个简单的任务，您需要的工具包应包含加速器、集成的开发环境(IDE)，以及一个声明式语言，使其对大部分组织是可消费和可实现的。毕竟，您毕竟，如果分析完全依赖于几乎不可能找到或学习的技能，您就无法在您的组织中普及大数据。除了文本数据是非结构化的这个事实以外，即使不考虑拼写错误、缩写或高级用法(如讽刺)，语言也是复杂的。因此，您需要足够深刻和灵活的系统来处理复杂性。

以下是该流程的一个示例，其中的文本分析应用程序读取一段文本，并根据各种规则派生出结构化信息。这些规则在提取器中定义，例如，它可以识别在文本字段中的实体名称。请看下面的文本:

In the 2012 UEFA European Football Championship, Spain continued their international success, beating Italy 4-0 in the Final. Spanish winger David Silva opened the scor- ing early in the game, beating Italian goalie Gianluigi Buffon. After a full 90 minutes of dominance, goalkeeper Iker Casillas accepted the championship trophy for Spain.

这些提取器的产品是一组注释的文本，如这段话中带下划线的文本。下面的结构化数据是从该示例文本中派生出来的:

Name	Position	Country
David Silva	Winger	Spain
Gianluigi Buffon	Goalkeeper	Italy
Iker Casillas	Goalkeeper	Spain

挑战是确保结果的准确性。准确性由两部分组成，查准率和查全率:

- **查准率** 正确性的一个衡量指标是结果集内的相关项目的百分比: “您获得的结果有效

吗？”举个例子，如果您想从UEFA锦标赛转录中提取所有与进球得分有关的每场比赛描述，而被识别为进球评论的76个段落中有30个与进球完全无关，您的查准率不到60%。总之，查准率描述了被识别的段落中有多少是被正确识别的。

- **查全率** 完整性的衡量指标，从文本中检索到的相关结果的百分比；换句话说，原始文本中全部有效字符串都出现了吗？例如，如果您想从UEFA锦标赛的视频中提取所有进球得分，并得到了人类专家发现的76段中的60段，那么您的查全率是约79%，因为您的应用程序错过了21%的进球得分。总之，查全率是被发现的匹配段落数在匹配段落总数中所占的比例。

分析师开发其提取器和应用程序，他们迭代地进行优化，以调优查准率和查全率。雪崩是一个很好的比喻。如果雪崩在滚落山坡时没有累积速度和更多的雪，它不会有太大的影响。提取器的开发其实是对提取器本身添加更多的规则和知识；总之，它的意义在于每次迭代都变得更强大。

我们发现，市场上大多数文本分析方法都对分析师提出了挑战，因为它们往往性能不佳(就准确度和速度两个方面而言)，并且它们难以构建或修改。这些方法让文本流过由提取器和过滤器组成的系统，没有优化。这种技术是缺乏灵活性且效率低下，往往造成多余的处理，因为在工作流后期被应用的提取器可能会执行一些之前已经完成的工作。从我们了解的情况来看，如今的文本工具包并不只是呆板和低效，它们的表达能力也是有限的(特别是，其查询可以支持的粒度)，这导致分析师必须开发自定义代码。然后，这会导致在优化结果集的准确性(查准率和查全率)时遇到更多延迟、复杂性和困难。

实现救援的注释查询语言!

为了满足这些挑战，IBM大数据平台提供了Advanced Text Analytics Toolkit，专门针对处理大数据中固有的挑战而设计。该工具包(最初的代码命名为SystemT)自2004年以来一直在不断开发中，其引擎随多个IBM产品提供，包括Lotus Notes、Cognos Consumer Insight等。IBM InfoSphere BigInsights (BigInsights)和IBM InfoSphere Streams (Streams)有了新的突破，它在Advanced Text Analytics Toolkit(及关联的加速器)中包括了SystemT技术，开放曾经是“黑盒子”的这一技术的自定义，并且使它与作为在产品中的功能交付时相比更通用。该工具包中含有一个声明性语言——注释查询语言(Annotated Query Language, AQL)，以及相关联的基于成本的优化器、一个用于编写规则的IDE、一个文本分析处理引擎(MapReduce和流式传输数据设置就绪)，和一些内置的文本提取器，它自带了通过在无数行业中的IBM客户协议预编制的数百条规则。Advanced Text

Analytics Toolkit还包含多国语言支持，包括对双字节字符语言的支持(通过Unicode实现)。提供一个优化器、一个AQL协助框架和调试工具，您可以看到Advanced Text Analytics Toolkit已准备好普及以SQL执行数据库查询的相同方式来对非结构化数据执行分析的能力。

Advanced Text Analytics Toolkit的特殊之处在于文本提取：为了确保高准确度(查准率)和全面覆盖(查全率)，该解决方案建立了许多具体的规则。其概念被内置到AQL和它的运行时引擎中，形成Advanced Text Analytics Toolkit的核心。AQL让您能够汇总这许多规则来代表每个提取器。例如，电话号码提取器可以包含数百条规则来匹配世界各地的人们表达这个概念的许多方式。此外，AQL是一个完全声明性语言，这意味着所有这些重叠的规则会被提炼和优化成一个高效的访问路径(类似于关系型数据库的SQL编译器，这是IBM研究人员最被开发此声明性概念的地方)，而从最终用户抽象出底层系统的复杂性。很简单，当您用AQL编写提取逻辑时，您告诉IBM大数据平台要提取什么，平台就会弄清楚如何提取它。这是IBM大数据平台的一个重要优势。使用声明性语言(例如，AQL、Streams Processing Language和机器学习统计语言)不仅在分析代码得到优化时具有显著的性能优势，并且它也对分析师隐藏了Hadoop的复杂性。请注意，我们引用的一些大数据声明性语言是IBM平台的一部分吗？尽管本章着重于文本分析，但认识到不同的大数据项目需要不同类型的优化，这一点很重要。例如，文本分析在很大程度上依赖于CPU进行处理。与此同时，在Hadoop咀嚼数万亿个键值对将消耗系统的I/O功能(如Hadoop相应的terasort和grep基准所示)。通过为手头特定的任务提供高度优化的优化运行时，大数据从业人员可以专注于分析和发现，而不是性能优化。

据我们所知，在当今市场上并没有其他完全声明的文本分析语言。您会找到高级和中级的声明性语言，但它们都利用无法自定义的锁定的“黑盒子”模块，限制了灵活性，并使其难以进行性能优化。

能够让文本提取器进化是非常重要的，因为在涉及分析时，极少东西是永远相同的。我们在社交媒体分析中经常看到这一点，流行的俚语术语或缩写词迅速变得“疲劳”(例如，几年前，很多人如果喜欢某个东西，他们会说“that′s sick!”但是，现在已经没有使用得这么频繁——我们对此感到高兴，原因很明显)。

AQL的设计目标是易于修改，当您进行修改时，新的代码被优化，以配合现有的代码。此外，AQL是针对重用而设计的，使您能够跨组织共享分析。您可以构建不同的提取器集，并将它们用作构建块，这样您就不必总是“从头开始”。

Advanced Text Analytics Toolkit包括内置的提取器集，适用于文本集合中常见的元素。例如，Person(姓名)、PhoneNumber、Address和URL是一部分随BigInsights和Streams提供的提取器。除了通用的提取器之外，还有大量面向社交媒体文本和常见日志数据类型的提取器集合。这些内置的提取器真正拉平了有效的文本分析开发的时间曲线。您也可以利用自己的自定义继续构建这些库，或在您构建的其他提取器中使用它们。我们将在第9章中介绍所有IBM大数据平台加速器。

如果您需要的提取器超出了开箱即用的范围，AQL是一个类似于SQL的语言，可用于构建新的提取器。它的表达能力很高，并且具有灵活性，同时提供熟悉的语法。例如，以下的AQL代码扩展现有的电话号码和姓名提取器，以定义一个新的提取器，专用于提取与某个特定的人相关联的电话号码。

```
create view PersonPhone as select P.name as
person, N.number as phone
from Person P, Phone PN, Sentence S where
Follows(P.name. PN.number, 0, 30)
and Contains(S.sentence, P.name) and
Contains(S.sentence, PN.number)
and ContainsRegex(/\b(phone|at)\b/,
SpanBetween(P.name, PN.number));
```

图8-1显示了前面的代码块中所定义的提取器的可视化表示。

当结合BigInsights和Streams的速度及企业级稳定性时，Advanced Text Analytics Toolkit代表了一个无与伦比的价值主张。与BigInsights和Streams集成的细节(详见图8-2)对于文本分析开发人员是透明的。在完成的AQL被编译和自动进行性能优化后，其结果是一个分析运算符图形(AOG)文件。对于BigInsights，可以通过BigInsights Web Console将此AOG提交为一个分析作业。在被提交之后，此AOG被分发到将在BigInsights集群上被执行的每个映射器。当作业开始时，每个映射器执行代码来实例化它自己的Advanced Text Analytics Toolkit运行时，并应用AOG文件。来自每个映射器的文件分割的文本在该工具包的运行时中运行，并且传回一个注释文档流作为结果集。

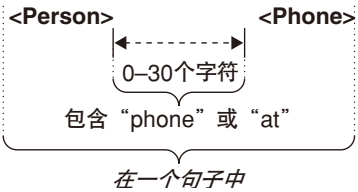


图8-1 代码示例中的提取器规则的可视化表示

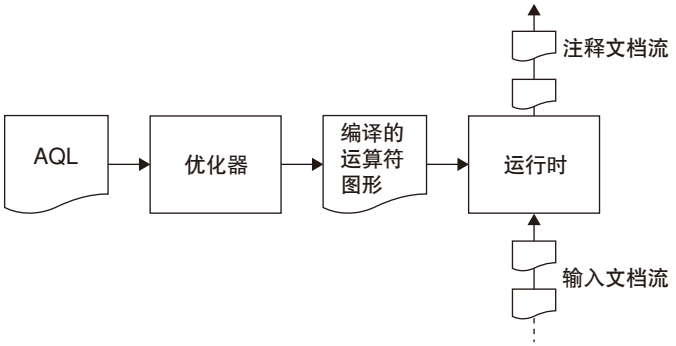


图8-2 使用Advanced Text Analytics Toolkit构建的分析的运行时流程

对于Streams，AOG被包括在Streams运算符中。在一个Streams节点上执行的过程中，运算符在该工具包的运行时传递流式传输的文本，运行时将结果集返回给运算符。

改变一切的生产力工具

Advanced Text Analytics Toolkit包括一组Eclipse插件，以提高您的工作效率。在编写AQL代码时，编辑器提供了自动完成辅助、语法高亮、设计时验证(语法错误的自动检测)等等，如图8-3所示。

文本分析其中一个最困难的方面就是入门。为了使它更容易，Advanced Text Analytics Toolkit包括一个工作流助手，使您能够选择自己知道会感兴趣的文本元素，并且它为您建立规则(见图8-4)。您可以为自己在使用的提取器选择文本的其他变化，以不断地改进这些规则。

另外还包括一个针对目标数据样品测试提取器的设施。构建文本提取器是一个多次迭代的过程，AQL工具的设计目的不仅是在分析师调整规则及其结果集时提供支持，还旨在促进开发人员和业务用户之间的协作。

分析师所面临的主要挑战是确定已应用到文本的变更的沿袭。可能很难辨别哪些提取器和哪些独立规则需要被调整，以便对结果进行微调。为了提供这方面的帮助，起源查看器(如图8-5所示)带有一个交互式的可视化功能，它显示究竟有哪些规则会影响生成的注释。

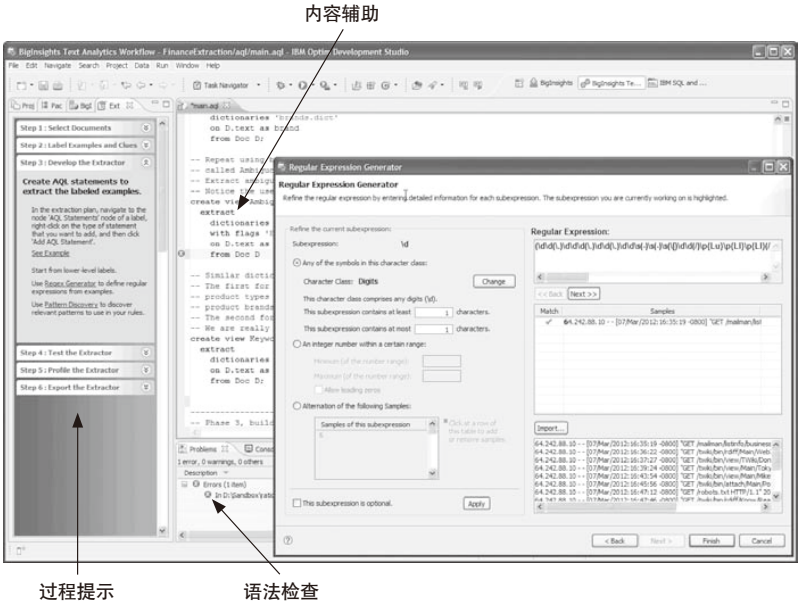


图8-3 一个插件为创建、调试和执行AQL提供了一个快速应用程序开发平台。

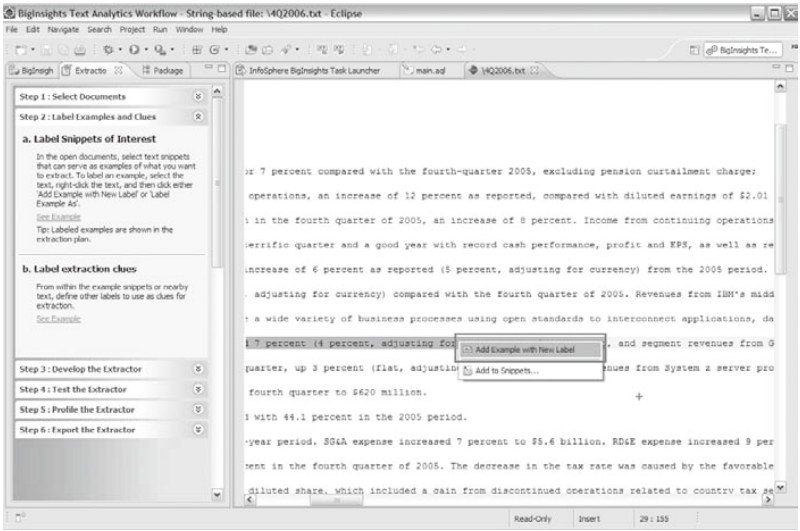


图8-4 AQL工作流助手: 文本分析入门的指南性帮助

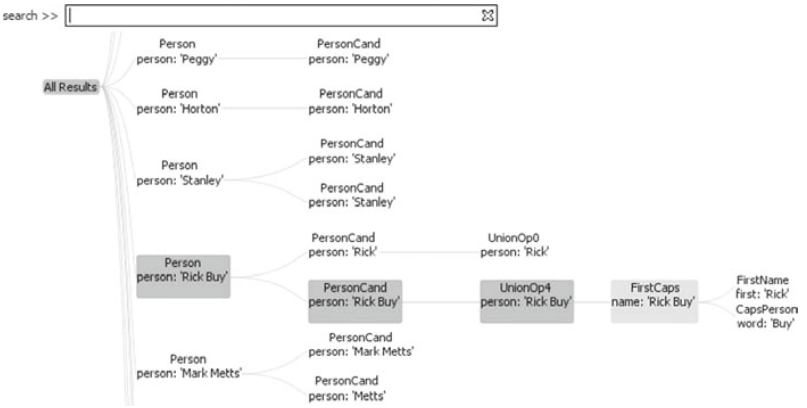


图8-5 起源查看器是Advanced Text Analytics Toolkit中所包含的多个开发工具之一。

想象一下，手动创建一个包含900条规则的提取器(如IBM大数据平台中开箱即用的某些提取器)；如果您犯了一个错误，您如何找出该错误来自哪里？起源查看器在此场景中是必要的。还有许多其他开发加速器特性，比如用于前期模式检测的IBM Many Eyes可视化引擎，等等。

总结

BigInsights Advanced Text Analytics Toolkit为您提供快速开发文本分析应用程序所需要的一切，它将帮助您从极大量的文本数据中获取价值。它不仅有广泛的工具支持大规模的文本分析开发，并且所生成的代码也是经过高度优化的，对运动中的数据和静止数据均易于部署。该工具包还包括一个丰富的提取器库，您可以定制和扩展该库。

9 IBM大数据分析加速器

IBM大数据平台包含多个分析大数据的通用工具。虽然这些工具的功能非常丰富，但为了货币化您的大数据资产，您不可避免地需要编写适合您的业务需求的应用程序逻辑。这和存储在关系型数据库中的数据没有任何区别，同样需要应用程序使数据库适用于特定的用例(例如，CRM软件)。当然，我们将注意到，关系型数据库行业已非常成熟。可以买数千种数据库应用程序，并且开发技能非常普及。如今，不容易找到开箱即用的大数据应用程序，组织只好采用自己部署(RYO)的方法。现在考虑一下，大数据开发技能由于稀缺而价格昂贵，我们打赌您正开始看到挑战.....和需求。

为了减少从大数据中提取商业价值所花费的时间，IBM已经开发了一套大数据分析加速器——向IBM大数据平台提供特定分析功能的软件模块。IBM的大数据分析加速器是从无数的客户协议学习到的专家模式的最终成果。您可以在各个行业使用大部分这些加速器，从大数据中提取真正的价值。目前已提供许多这些加速器，并且还有更多正在开发中，将来会发布。在本章中，我们将介绍三个目前可用的加速器: Machine Data Accelerator(机器数据加速器)、Social Data Accelerator(社交数据加速器)和Telco Data Accelerator(电信数据加速器)。

IBM Accelerator for Machine Data Analytics

如今的企业在很大程度上依赖于其IT基础架构的正常运行时间。我们的一个客户通过与股票行情自动收录器类似的应用程序来赚钱，该应用程序传递值得关注的新闻项目，如果它不运行，他们就无法赚钱。另一个在线零售商估计，在处理在线订单的过程中会产生17种日志。Web日志、应用程序服务器日志、Hibernate(休眠)日志、数据库日志、驱动程序日志，等等，都包含将事务作为一个整体的不同视图。这些客户告诉我们，他们无法承受停机时间，所以他们在所关注的领域中投入了大量的资金和智慧。我们认为，虽然关于保持系统的高可用和灾难恢复已经有很多想法和规划，但在涉及我们的IT环境的高度互联性质(每个人的设备都是联网的，还有许多系统通过这个线缆迷宫输送信息)时，我们仍然太想当然了。当然，在出问题的时候，我们对这些系统的依赖性明显得让人痛苦。如果没有连接性这一命脉，整个部门的活动都可能被迫进入停顿状态。因此，当IT故障发生时，有一个要恢复连接的极端紧迫感。问题是这样的: 如果正常运行时间如此关键，为什么我们只有极少人会制作日志信息的集合，并发现日志事件之间的趋势线相关性？有些日志事件孤立而言似乎

无害，但结合其他IT活动的上下文就会被证明是引发下游中断的根本原因。大数据技术为企业带来了前所未有的机会，可以创造针对无数的日志文件的洞察，以产生有关过去的错误以及如何防止事情在将来离开正轨的提示和线索。

正如我们在前面的段落中所提到的，有两个主要因素带来了如今的IT中断挑战：IT系统之间高度的互联性和不断增长的相互依赖关系，以及对这些系统的大量使用。诊断故障的根本原因的关键在于系统管理员梳理来自其多个服务器的IT日志(也被称为机器数据)，并找出导致中断的事件链的起源。正是这些日志的性质构成了最大的挑战。所有这些不同的系统都在不同的位置存储日志，使用不同的文件格式，并使用不同的元素显示样式，如日期时间信息。极大的数量甚至使此信息的种类更加难以处理。组织堆积许多个TB(数万亿行)的机器数据是常见的情况。今天，许多组织只是清除这些数据，几乎就像它是无用的副产品一样(这就是为什么我们通常把它称为数据废气)。显然，能够在系统日志中找到隐藏的价值就是一个大数据挑战。

为了迎接这一挑战，InfoSphere BigInsights (BigInsights)自带了IBM Accelerator for Machine Data Analytics(非正式地称为Machine Data Accelerator，或缩写为MDA)，这是一个特殊的模块，旨在处理日志数据分析的整个生命周期(如图9-1所示)。

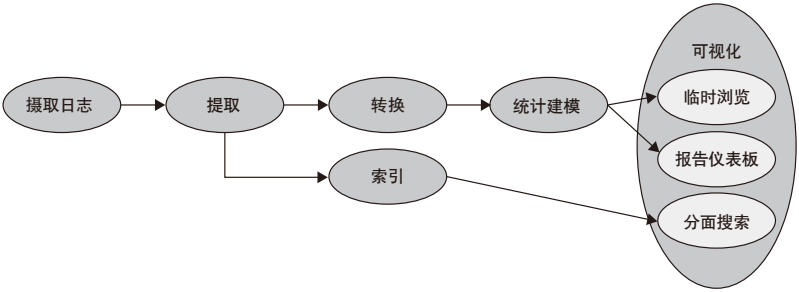


图9-1 机器数据分析的生命周期

摄取机器数据

机器数据分析生命周期的第一阶段是从IT系统将日志摄取到HDFS(HDFS的解释请参见第5章)。MDA包括一个摄取应用程序，它处理这种数据移动操作，而且还帮助为将在BigInsights中发生的后续数据处理准备机器数据。

MDA的数据摄取函数以批量形式接受日志，其中每个批处理代表一种日志类型。除了日志数据本身，每个批处理还包括一个元数据文件，它描述关键特征和固有的假设。MDA必须使用此信息才可以正确地解析和规范化数据。机器数据中的一个共同特点是，关键的元数据元素(如年份或服务器名称)要被编码为存储日志的文件的名称。显然，当来自不同的时间段和不同系统的日志被放在一起时，该信息需要被考虑进去。

提取

当机器数据被存储在HDFS中之后，您可以利用IBM大数据平台的文本分析功能来解析它，并提取感兴趣的项目。MDA的提取函数的目的是把来自多种来源的日志记录转换为一致的格式。如果不这样做，就几乎不可能执行有意义的分析，因为这么多日志格式有着明显的差异。

MDA包括对以下日志类型的开箱即用的支持: Cisco syslog、WebSphere Application Server、Datapower、Webaccess和包含头部的CSV。MDA还包括了一组常见于所有日志的基本提取器(例如，IP地址和日期时间)，它们可用于其他日志类型。对于需要更多定制的情况，您可以使用Advanced Text Analytics Toolkit(在第8章中介绍)并定制现有的提取器，或者构建自己的提取器。例如，您可以打开MDA的Cisco weblog提取器，并对其进行修改，以满足您的日志类型。其实，您可以将MDA组件视为模板，并且以您想要的任何方式自由定制它——让您更快地获得在数据中的价值。

MDA的提取函数解析一个批处理中的所有日志记录(使用适合于所应用的日志类型的提取器)，并以能够与来自其他系统和时间段的日志一起进行分析的形式写入它们。这个过程涉及以下步骤:

1. **记录分割** 这涉及到解析日志并识别日志记录边界。来自某些源的机器数据(例如，Cisco syslogs)将记录表示为单行。来自另一些源的数据(例如，WebSphere Application Server)将记录表示为多行。

2. **字段提取** 在日志记录被分割后，根据与当前批处理的日志数据关联的日志类型的规则提取每个字段。

3. **事件标准化** 当分析来自不同设备、软件和应用程序的机器数据时，需要以一致的方式存储记录内的时间戳，要考虑到不同的时间戳格式、时区或数据本身缺少的与时

间戳相关的其它信息。必需这样做的原因是，为了减少的日志大小，来自每个记录的信息经常被省略。例如，在每个记录中的时间戳可能会缺少年份或时区，它已在文件名中提供，或在外部已知。MDA使用在批处理的元数据中提供的值来填充缺少的字段。如果没有这种时间戳数据的标准化，就不可能公平地比较日志文件。

4. **事件丰富化** 用户指定的元数据(如服务器名称、数据中心名称或应用程序名称)在摄取过程可以与机器数据批次关联。例如，与服务器本身直接相关的日志记录批次通常不包括服务器名称，但与来自其他服务器的日志批次一起分析这些信息时，它将是非常有用的。

5. **事件一般化** 机器数据记录通常包含不同的值，如时间戳、IP地址、测量值、比例和消息。通过用恒定值(屏蔽)替换不同的值，事件可以被一般化。一般化的事件被收集和赋予唯一的ID，然后将其用于下游分析。这些一般化的事件，可以在频繁序列分析中使用，以确定哪些一般化事件序列发生的频率最高。它们也可以被用于显著性测试，以识别哪些一般化事件相对于某个特定的错误而言是最显著的。被屏蔽的字段因日志类型而不同。事件一般化是可选的，若用户不提供任何字段，则不执行一般化。

6. **在BigSheets中的提取验证** 在运行提取操作之前，您可以在BigSheets中预览结果，以确保提取了正确的字段，并且正确地应用了标准化、丰富化和一般化操作。

7. **提取的日志存储** 将得到的数据被存储为在目录层次结构中的压缩二进制文件，其中的每个目录都包含的，在来自一个日志批次的已解析的日志记录。日志被格式化为JSON记录；每个记录都包含原始日志记录和为日志记录提取的字段。

索引

为了方便对机器数据进行搜索，提取的日志必须被索引。搜索界面也支持分面浏览(通过选择感兴趣的类别，可以钻取一组日志记录)。要启用此功能，您可以自定义要使用哪些分面，以及它们的标签应该是什么。

转换

提取的机器数据代表许多单独的日志记录，其格式可以与来自其他时间段和来源的日志记录一起分析。其中一个这样的分析以会话化(sessionization)的概念为基础，也就是说，将对应于一个时间段或基本活动的日志记录分为一组(这是通常在基于Web的购物车分析中使用，但它对于网络分析也同样有用)。通过会话的镜头分析机器数据的能力对于根本原因分析、模式识别和预测建模都是有用的。MDA中的转换函数执行两种类型的会话化: 时间和事件上下文关系(您可以在这里联接两个不同的日志集)。

要利用的时间会话化，使用MDA基于时间片将日志记录分组为会话。MDA将根据所提供的分区键(例如，一个机器ID或进程ID)划分记录。一旦记录被划分，您只需指定一个时间间隔阈值，MDA将组合日志条目，直至达到时间间隔阈值，从而为每个分区将日志记录拆分为会话。

事件上下文关系会话化使您可以围绕来自一个日志集合的特定事件类型，以及来自另一个日志集合的相关信息定义会话。要开始该处理，首先要识别需要分析的事件类型——这被称为种子事件，MDA将围绕它构建会话。其次，识别需要为起源事件提供上下文的日志信息类型，并为“事件”日志类型和“上下文”日志类型指定一个分区键。(分区键是事件日志记录和上下文日志记录中均存在的一组字段)。利用该信息，MDA将记录拆分为由您指定的事件(如断开的连接)所定义的会话。

统计建模

现在，您的大量不同格式的机器数据已经过标准化和转换，您可以开始执行一些有意义的分析。即使不考虑不同的日志类型，数量也足以让统计分析令人望而却步。在Hadoop上下文中，即使有可能高效地处理庞大的数据量和多样性，也难以编写统计算法的程序。BigInsights在这里可以让您更轻松，因为它有一个工具包，可以实现机器学习和深度统计分析。MDA利用部分这些统计算法，帮助揭示锁定在机器数据中的宝贵信息。目前有两种统计模型可供MDA使用: Frequent Subsequence Identification和Significance Analysis。(我们的律师不喜欢我们暗示未来，但是想象一下有更多的统计模型任您支配的MDA——这是迟早的事。)使用BigSheets可以方便地对这两个统计模型提供的输出进行可视化和图表化。

Frequent Subsequence Identification模型显示哪些事件序列在不同的会话中发生得最频繁。使用该分析可以提示许多有趣的模式，这支持主动管理，以防止未来的问题。例如，您可以识别在故障条件之前频繁发生的事件序列。Significance Analysis帮助针对错误条件识别哪些事件和事件模式是最显著的。

可视化

MDA包括一些工具，使您能够以图形方式钻取机器数据并可视化之前隐藏的趋势。事实上，所有提取的数据和会话化的记录都可以供BigSheets中的临时查询和可视化进行处理。除了细化报表和查询之外，您还可以使用强大的分面搜索工具来探索机器数据。

分面搜索

机器数据被提取并索引后，您可以使用BigInsights中包含的Data Explorer图形化分面搜索界面来浏览它。这是搜索机器数据的一个快速方法，可以加快故障排除——您

可以定义一个小的时间窗口并对您的日志执行基于分面的向下钻取，从而将精力集中在一个非常大的记录集的一小部分上。图9-2显示了分面搜索界面，其搜索结果按时间范围和一个搜索字词进行过滤。在界面的左侧，您可以选择其他的类别，以进一步细化搜索结果。

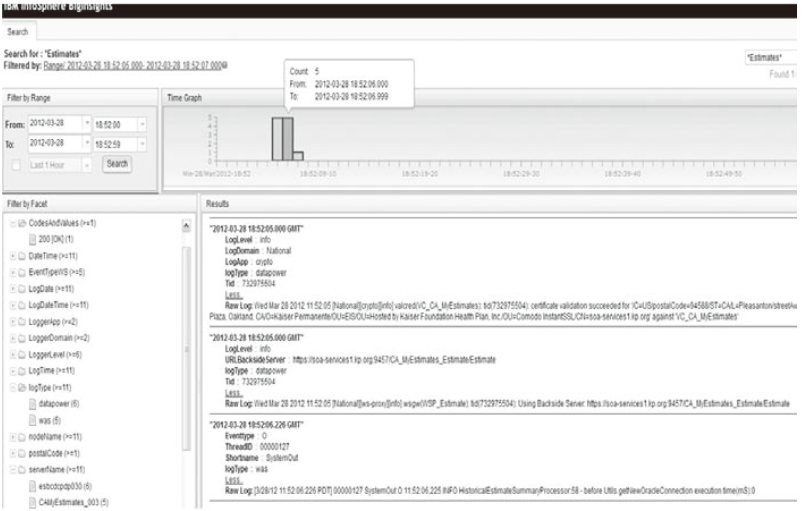


图9-2分面搜索界面

IBM Accelerator for Social Data Analytics

所有类型的大数据都没有像社交媒体一样被热炒和激烈辩论。营销人员对可以了解在Twitter上关于自己的品牌有何说法这个前景垂涎三尺。正因为如此巨大的需求，围绕社交媒体消息的解读已经涌现出一个小作坊行业。在几乎每一个重大的文化活动中，我们都以Twitter情绪图表的形式看到这一点。例如，在2012年夏季奥运会，点亮伦敦眼(在伦敦滨水区的巨型摩天轮)的灯光颜色反映了人们对这个赛事的整体情绪。

几乎无一例外地，来自数量庞大的社交媒体消息的情绪表示是很肤浅的。我们发现，大多数表示由一组简单的规则构成，将正权重分配给正面的单词，将负权重分配给负面的单词。这对于一些推文也许是可行的，但不能认为它是可靠的。例如，请考虑这样的句子:

"The amazing number of lip synched performances during the 2012 closing ceremonies inspired me to take a walk and have my own lip performance at the local pub with a pint."

即使两个带下划线的单词被认为是正面的，这篇推文却是极度负面的。这句话证明，上下文是非常重要的，自由格式的文本的任何认真分析都必须通过更复杂的方法来推导出意义，而不是一个简单的词汇分类。在本书的一些章节中，我们已经指出了这一点，而这正是IBM大数据平台的其中一个优势。

此外，情绪本身并不足够。如果您在一家汽车公司的市场营销部门工作，您想知道某人是否有兴趣购买贵公司的产品。更好的是，您可能希望看到市场细分的详细信息，那么您就可以对人们如何响应您的营销活动进行细分。

在BigInsights、Streams和Advanced Text Analytics Toolkit的功能的基础上，IBM开发了Social Data Accelerator(缩写为SDA，但正式名称为IBM Accelerator for Social Data Analytics)，它提供了一套丰富的文本分析规则，让您可以准确和精确地了解人们在网上说些什么。

SDA包括专注于多个行业的线索生成和品牌管理的提取器。线索生成的重点是发现潜在客户，而品牌管理则是人们对品牌以及任何竞争对手有何感受。此外，针对特定用例的每一组提取器被划分为确定一个社交媒体消息的意义的规则，以及有助于构建一个特定用户的档案信息的规则。为了将这些提取器的输出集中在一起，SDA包括了一个工作流，可以帮助您在从摄取一直到可视化的整个过程中管理流过这些提取器的数据流。

反馈提取器: 人们在说什么?

正如我们刚刚看到的，要了解人们的社交媒体消息背后的意义，仅仅对正面和负面的单词进行评分是不够的。原始文本(如推文)的语义理解是复杂的，并且要求仔细的分析。这是出于这个原因，IBM已经在SDA中为情绪分析工具构建了行业特定的提取器。这些反馈提取器代表应用到您所提供的品牌、产品或服务的行业特定的规则集(适用于零售、金融和娱乐)。以下是SDA规则寻找的反馈类型。

- Buzz 讨论的量
- Sentiment 满意或不满意的程度。此粒度可以相当细。例如，如果您发布推文表示自己喜欢电影Moneyball，但不喜欢男主角(这只是一个例子，我们认为Brad Pitt很酷)，SDA可以隔离对这部电影的情绪。
- Intent to buy 花钱的承诺水平
- CustomerOf 判断一个人是否是现有的客户(例如，有人说，他们买了一个特定的产品)

档案提取器: 这些人是谁?

联系人们在说什么的上下文，了解谁在表达情绪，这是非常有价值的。档案提取器通常在每次添加新批次时被应用在大型社交媒体消息集上。在最初几次运行这些提取器时，所获得的档案数据将是稀疏的，但随着社交媒体消息集合的增长，更多档案元素将被填充到要跟踪的更多用户。除了发布者的在线档案信息(如他们的用户名或Twitter档案)，档案提取器从社交媒体消息确定以下信息:

- Gender 发布者的性别，通常从发布者的在线档案信息决定
- Location 发布者住在哪里
- Parental status 发布者是否有子女
- Marital status 发布者是否已婚
- Employment status 发布者是否有工作

工作流: 把一切聚集在一起

为了成功地评估人们对您的品牌或产品有何言论，SDA包括了一个工作流(参见图9-3)，以协调分析社交数据需要采取的所有步骤。

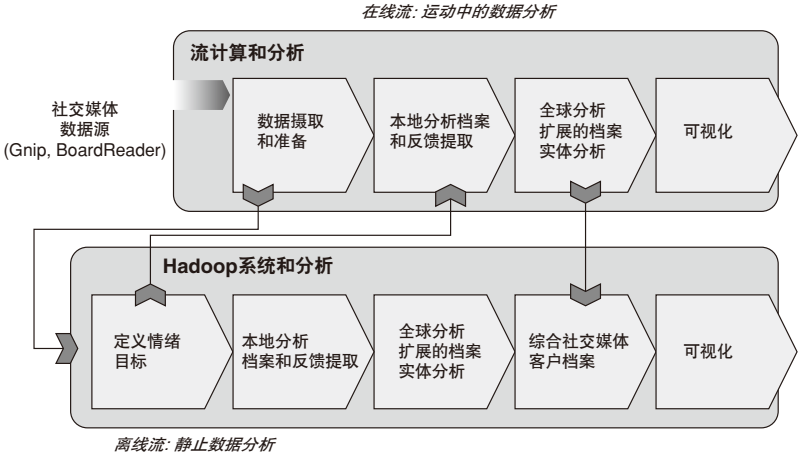


图9-3 社交数据分析的生命周期

摄取

社交媒体分析的第一个阶段是摄取社交媒体消息。在SDA中，这个阶段由Streams处理，它拥有GNIP服务(它提供来自Twitter的消息)和BoardReader服务(提供来自微博、博客、新闻组、论坛等的消息)的源运算符。虽然Streams将数据摄取到BigInsights中，但它也执行一些垃圾消息的初始过滤，以解决真实性的问题。当然，可能您会希望调查这些垃圾消息；也许其中一部分并不是垃

圾，您可以通过自定义来进化过滤过程，毕竟，它是一个学习环境。就像敏捷开发一样，迭代的次数越多，解决方案会变得越好。Streams还可以在数据被流式传输到组织中的时候提供对数据的实时洞察。

SDA是IBM在该领域数千次迭代的结果，但如您所知，旅程永远不会结束。

定义情绪目标

当我们这样写时，这似乎相当明显，但您需要在分析开始之前就告诉SDA您正在寻找什么。当涉及到文本提取时，您必须定义要提取的元素，并从那里开始发现。例如，您可以向SDA提供您有兴趣跟踪的一组品牌、产品和服务。您也可以定义产品类别信息和产品的别名，从而让分析变得更复杂。SDA摄取已定义的感兴趣术语，并把它转换成一个AQL字典(有关更多信息，请参阅第8章)，在SDA的下游处理中所执行的反馈提取器会考虑到它。

本地分析

该阶段是针对该批数据的两个文本分析通道中的第一个。在这一阶段，使用反馈提取器和档案提取器对每个文档进行单独分析。在分析文本本身之前，先执行额外的垃圾过滤，并隔离提到品牌、产品或服务的文本。此时，对预处理的社交媒体消息运行反馈提取器和档案提取器。档案提取器的输出被添加到存储在BigInsights中的全球用户档案。反馈提取器的输出被写入BigInsights，格式为逗号分隔值(CSV)格式，以利用BigSheets实现简单的可视化。

全球分析

更新的用户档案到位后，对这批数据执行第二轮分析，采用实体分析来确定各数据源有哪些用户是相同的。同样，这个阶段的输出作为更新被写入存储在BigInsights的全球用户档案。

流式传输模式

虽然SDA可以使用BigInsights以批处理模式处理社交媒体消息，但它也能够使用Streams以流式传输模式运行。在这种情况下，将情绪目标的定义传递给BigInsights，它生成AQL，所生成的可执行代码被传递给Streams。本地和全球分析以同样的方式在Streams上发生，它以批处理模式执行，但每个记录都在从社交媒体源适配器流进来时被处理。

可视化

SDA提供其分析结果的可视化和报告。您也可以在Data Extractor Application Builder仪表板中发布这些报告。图9-4显示了一个示例仪表板，在那里您可以显示整体情绪，同时强调每个客户的情绪。

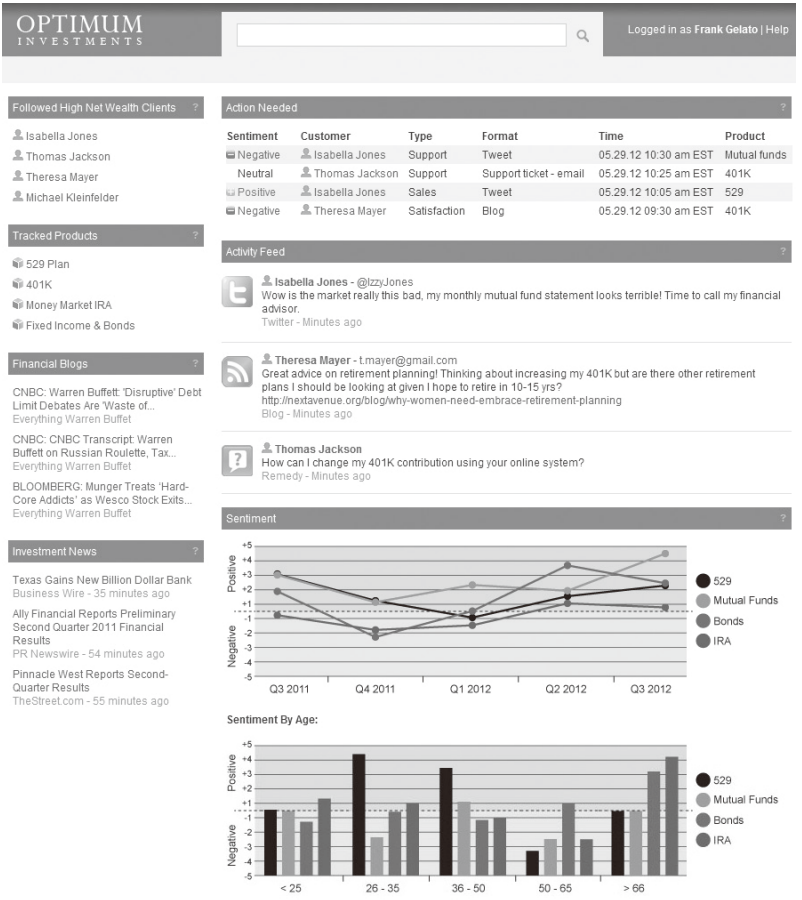

OPTIMUM
INVESTMENTS
Logged in as Frank Gelato | Help
Followed High Net Wealth Clients
Isabella Jones
Thomas Jackson
Theresa Mayer
Michael Kleinfelder
Tracked Products
529 Plan
401K
Money Market IRA
Fixed Income & Bonds
Financial Blogs
CNBC: Warren Buffett: 'Disruptive' Debt Limit Debates Are 'Waste of...
Everything Warren Buffet
CNBC: CNBC Transcript: Warren Buffett on Russian Roulette, Tax...
Everything Warren Buffet
BLOOMBERG: Munger Treats 'Hard-Core Addicts' as Wesco Stock Exits...
Everything Warren Buffet
Investment News
Texas Gains New Billion Dollar Bank
Business Wire - 35 minutes ago
Ally Financial Reports Preliminary Second Quarter 2011 Financial Results
PR Newswire - 54 minutes ago
Pinnacle West Reports Second-Quarter Results
TheStreet.com - 55 minutes ago
Action Needed
Sentiment Customer Type Format Time Product
Negative Isabella Jones Support Tweet 05.29.12 10:30 am EST Mutual funds
Neutral Thomas Jackson Support Support ticket - email 05.29.12 10:25 am EST 401K
Positive Isabella Jones Sales Tweet 05.29.12 10:05 am EST 529
Negative Theresa Mayer Satisfaction Blog 05.29.12 09:30 am EST 401K
Activity Feed
Isabella Jones - @IzzyJones
Wow is the market really this bad, my monthly mutual fund statement looks terrible! Time to call my financial advisor.
Twitter - Minutes ago
Theresa Mayer - t.mayer@gmail.com
Great advice on retirement planning! Thinking about increasing my 401K but are there other retirement plans I should be looking at given I hope to retire in 10-15 yrs?
http://nextavenue.org/blog/why-women-need-embrace-retirement-planning
Blog - Minutes ago
Thomas Jackson
How can I change my 401K contribution using your online system?
Remedy - Minutes ago
Sentiment
Positive
Negative
+5 +4 +3 +2 +1 -1 -2 -3 -4 -5
Q3 2011 Q4 2011 Q1 2012 Q2 2012 Q3 2012
529
Mutual Funds
Bonds
IRA
Sentiment By Age:
< 25 26 - 35 36 - 50 50 - 65 > 66

图9-4 社交数据仪表板

IBM Accelerator for Telecommunications Event Data Analytics

电信(telco)行业一直是IBM大数据平台的早期积极采用者，尤其是将它应用于处理运动中的数据。考虑到电信公司所面临的数据管理挑战，特别是在高速数据方面的挑战，这并不奇怪。根据为电信企业提供解决方案的经验，IBM已构建了TelcoDataAccelerator(缩写为TEDA，但正式名称为IBM Accelerator for Telecommunications Event Data Analytics)，它专注于运动中的呼叫详细记录(CDR)的处理。TEDA随Streams提供，我们已在第6章中讨论过它。

CDR处理对于如今的电信服务供应商是一个尖锐的难题，主要是因为庞大的数量和极高的速度。每当有人拨出电话，电信交换机就创建一个CDR，其中包括如主叫号码、被叫号码、开始通话的时间戳、通话类型(语音或SMS)、通话持续时间、质量信息等等信息。随着手机的出现，CDR日志的数量大幅增加，因为同一个移动通话每次被转移到另一个手机信号塔时都会创建额外的CDR。对于拥有众多客户的电信提供商，主动处理这些CDR日志是非常困难的，在高峰使用时段尤其如此。举个例子来说明这些数量的规模，我们的大型亚洲电信运营商客户每天必须处理约六十亿个CDR，其中每天的峰值速率是每秒超过200,000个CDR！

大多数电信提供商批量加载CDR数据到关系型数据库，然后执行CDR处理(包括转换和分析)，实时用于收入保证或欺诈检测等应用程序。也有后处理的活动，如删除重复条目。这意味着在许多小时的延迟之后，电信业务才可以跟踪收费活动或其他问题。

TEDA旨在将大部分CDR分析和后处理活动推出静止持久层，并使用Streams在生成CDR后立即在运动中完成工作。这种方法不仅让电信运营商能够更直接地访问计费信息，还可以通过对CDR数据执行运动中的分析而收获丰富的附加信息。

在TEDA中的技术已在我们刚才提到的亚洲电信运营商客户以及其他客户处成功实施。TEDA可提供的拐点是惊人的。我们的亚洲电信运营商客户以前要用12个小时来处理它的CDR数据。现在，从生成CDR开始计算，对CDR执行基于Streams的摄取和转换只需不到一分钟就可以完成，这使他们能够获得实时的洞察并提供新的服务，如基于使用的优惠。另一个好处是处理效率，与电信运营商目前所使用的典型批处理应用程序所要求的处理或存储量相比，Streams通常也只需要其十分之一。

以下各节介绍TEDA中包括的三个主要分析功能：CDR丰富化和重复数据删除、网络质量监测，以及关键客户体验指标跟踪。

呼叫详细记录丰富化

使用TEDA，Streams可以摄取CDR并近实时地执行丰富化和重复数据删除活动。图9-5显示了TEDA架构的一个高层次视图，包括TEDA逻辑的流程图和它们与电信公司的事务处理系统中的各种数据存储的交互。

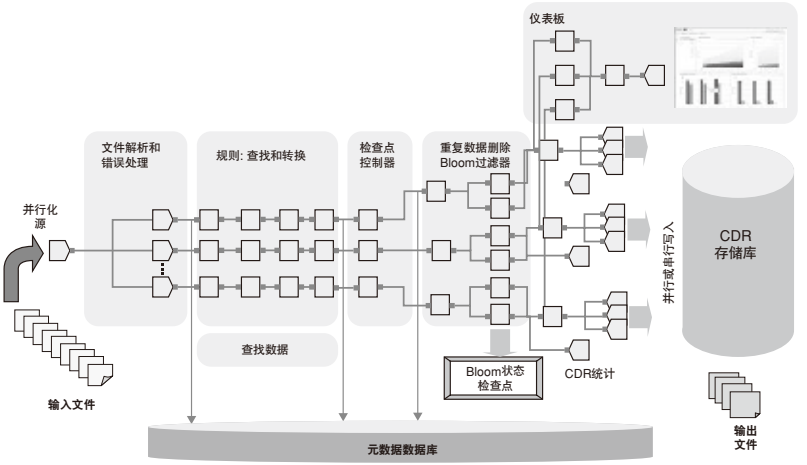


图9-5 TEDA架构

在许多司法管辖区中都有很多关于CDR数据的监管要求。因此，事务处理系统需要详细跟踪CDR何时被成功处理并写入CDR数据存储。Streams TEDA 应用程序持续监测它的目录以发现新的CDR，并写入记录，以维护每个文件的状态。一个并行化运算符将CDR分割成多个并行分支(或路径)，以加快处理并将完整的并行化技术应用到流。

TEDA支持Abstract Syntax Notation One (ASN.1) CDR格式和其他专用格式。对于其他CDR格式，TEDA摄取易于定制的规则，以适应这些格式的变化。

TEDA包括700多条规则，它们代表由沉浸在电信中的IBM人创建的专家模式，以促进CDR分析流程。TEDA还包括一系列在内存中查找和表查找，使用客户的重要性、客户ID，以及通话的预计收入等信息丰富CDR。

Streams在CDR丰富化的过程中也执行重复数据删除。电信交换机总是为每个CDR创建两个副本，以防止数据丢失，但是必须删除重复，以确保客户被结算两次。TEDA采用Bloom Filter算法来消除重复，它优化了性能和内存消耗。由于存在交换机故障的可能性，重复的CDR可能在长达15天后才会出

现。这意味着，每个CDR都必须与15天的数据(可能是数十亿个CDR)进行比较。通常情况下，该处理在CDR数据仓库中完成。然而，利用TEDA，它现在可以与CDR分析同时完成，这减少了仓库中的工作负载，使仓库可以专注于分析和报告应用程序，而不是删除重复的记录。

最后一组运算符将CDR写入CDR存储库(CDR仍然需要被储存在这里，基于多个原因，如合规性、数据治理、洞察发现，等等)。在TEDA应用程序接收到CDR已经被写入存储库中的确认后，控制信息被发送回源运算符，以更新CDR状态信息并删除相关的输入文件。

网络质量监测

TEDA的核心功能在于CDR丰富化和重复数据删除，传统上这是通过批处理使用静止技术完成的。但是，在摄取时立即处理CDR可以产生一些有趣的可能性，以便从CDR数据获得更多价值。具体来说，CDR中有些数据有助于网络质量监测。随着CDR流过图表，TEDA对完成的和掉线的通话进行计数，同时记录上下文数据，如供应商、国家、地区、手机ID、终止代码和手机站点。TEDA利用这些数据，提供如图9-6中所示的仪表板，它显示了按地区排序的CDR的当前吞吐量。历史汇总被存储在数据库表中，用于将来在Cognos中的可视化。

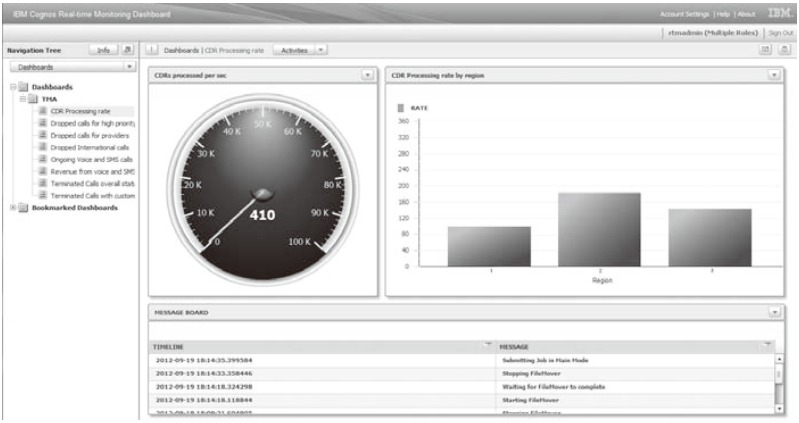


图9-6 CDR指标仪表板

客户体验指标

TEDA的另一个好处是能够跟踪客户体验的关键指标，最重要的指标是掉线的通话。使用一些由TEDA添加的丰富化的CDR数据，您可以查看最重要的客户的摘要。图9-7显示了一个有关高优先级客户所经历的掉线通话的仪表板报告。

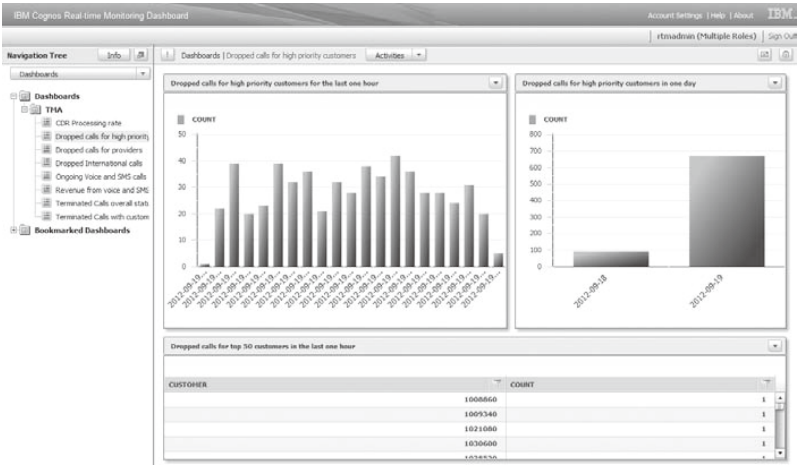


图9-7仪表板报告高优先级客户所经历的掉线通话

类似于网络质量报告，此客户体验数据的历史汇总被存储在数据库表中，用于将来在Cognos中的可视化。

总结:
为您的生产力加速

在所有三个加速器中的共同点是: IBM提供工具和一个经过优化的工作流，以减少从大数据中获取可操作信息所需的时间。利用Streams和BigInsights，您就拥有两个强大的引擎，可以处理广泛的运动中数据和静止数据。若配合使用这些分析加速器，处理能力就可以与开箱即用的分析功能相结合，可以大大加快您的部署。利用IBM大数据平台，不仅让您变得更智慧，还会变得更快！利用一组加速器，IBM将专业知识和经过验证的使用模式一起打包在一个机箱中，使得货币化和分析大数据的速度比以往任何时候都更快。

大数据世界中的集成和治理

10 治理还是不治理: 大数据世界中的治理问题

大数据是否需要治理渐渐成为一个大数据的热门话题(我们一开始就应该讨论这个问题)。大数据是一种现象，它正在不断地改变每一个系统的数据特征，为了让大数据可用，需要对大数据进行治理，使其更确定、更可信。有些人认为，需要以原始形式(保持保真度)来分析大数据，任何形式的治理或“清理”企图实际上可能会丢掉一些有价值的内容。其他人则认为，如果需要，治理功能可以简单地“内置”在大数据的生态系统中。这两种观点都是错误的。回答本章的标题提出的问题，答案是肯定的: 大数据需要治理。

想到这个问题时，道理显而易见。如果保存在存储库中的数据要经受治理，事实就是要引入更多的数据，或不同类型的互补性数据，或持久化引擎时，并没有改变任何东西。假设您的客户在Facebook页面上和您成为朋友，并同意与您分享他们的信息，如果他们以后与您的企业解除好友关系，根据Facebook的条款和条件，您将不再允许使用该信息。这构建到您的社交情感治理流程中了吗？如果您使用的是测试环境，其数据包含个人身份信息(PII)，无论是将其包装成漂亮又整洁的关系型架构，强调参照完整性规则，还是基于一致性机制的文件系统，好比狂野的西部，它仍然是PII数据。观点是，数据就是数据，如果今天要进行治理，在用不同类型的数据或更好的保真度来强化洞察时，为什么数据要改变？

我们时刻提醒客户，治理一个信息管理平台有多么重要。毫无疑问，大数据平台的选择应该包括一些治理的讨论，这是本章的重点。

为什么要治理大数据？

让我们从广义上理解治理来开始本节的内容。治理是关于如何管理数据的一整套策略。策略可能(也可能不会)包含主动治理数据(清洁，保护等)，但所有数据都应该有一个策略，体现出是否要治理数据的决定。没有人会主张在不了解情况时做出如何处理数据的决定，但可悲的现实是，有很多人采用这种方法。这将他们置于不幸的“无功治理”困境，在问题出现时去解决问题(或更糟糕的违规行为)，这是更昂贵和复杂的办法。因此，组织必须跟踪所有的数据，并定义如何管理数据的策略。让我们看一些例子。

考虑一下围绕数据的生命周期和管理实践。从数据进入企业的那一刻起，它就有一个过期日。在某些情况下，虽然数据可能永远不会过期，但在不经常访问数据时，它必然会有一个冷却期。然而此处有一个问题: 您真的知道数据的过期日和老化策略吗？您在明确的策略下管理数据吗？您知道什么时候可以归档数据？什么时候可以合法地删除数据？例如，药物临床试验需要在试验对象去世后，保留的时间要超过10年，大多数财务记录必须保留7年，Facebook已经同意Facebook帖子保持20年的时间，以供联邦隐私审计之用。

现在，再考虑一下保留问题在大数据世界里可以被放大多少。大数据往往有一个短暂的保质期，并且可以快速累积。

如果不定义生命周期策略，并用自动化技术加强这些策略，则会被所积累的大数据压倒，或着，您的管理员将花费大量的手工劳动来决定如何以及何时删除数据或让数据退役。考虑到社交媒体数据时，您需要让这些数据保持多久？因这些数据而得到的洞察呢，它们是怎样联系在一起的？同时，一些大数据承诺保持语料库信息(完整的历史)，因此，尽管您可能永远不会删除这些数据，但它仍有其热度。

许多大数据用例涉及到分析敏感信息。组织必须定义安全策略来维护这些信息，并且还必须监督和执行这些策略。

大数据的完整性方面是非常热门的一个话题，它有着一个性感的术语，真实性，我们在第1章中介绍过。您需要确定是否采用与传统数据同样的方法来清理大数据，或者，您是否愿意承担清洁数据的风险，如可能会失去潜在的、有价值的洞察。答案完全取决于您打算用这些数据做什么。在某些用例中(如客户分析)，需要或只能受益于更高质量的数据。在其他用途的情况下(如欺骗性身份分析)，可能需要与输入时完全相同的数据，以便发现虚假身份模式，而不是依赖于更高质量的数据。

许多大数据用例以关键的主数据管理(master data management，MDM)概念为中心，如客户、产品、地点和供应商。但是，许多企业在大数据的热潮来临前，还没有为这些领域建立一个单一版本的事实。考虑一个以社交媒体为基础的客户分析应用程序。其主要的出发点之一是了解客户。其中的联系是，许多MDM项目的目标是提供一个单一的客户视图。MDM和大数据之间的联系以组织自身最关注的、最有价值的业务实体为中心，包括客户、产品和资产。这是MDM和大数据之间的联系。MDM是许多大数据用例的一个不错起点，它还提供了一个逻辑中心来存储从大数据分析收集到的洞察。例如，如果考虑一个以人员为中心的主数据项目，然后从Twitter或Facebook上提取生活事件，如关系状态的变化，宣布婴儿出生等，并丰富该主信息，同时将其作为一

种系统的信息源。MDM可在大数据治理措施中起到举足轻重的作用，它提供了一个已治理的、单一版本的事实，通过在大数据措施中得到的有价值洞察来进行说明。

我们以前注意到，许多组织不承认有必要治理大量的新数据，这些数据是其大数据措施带到企业中的，因此，在大数据项目的规划阶段没有正确地考虑数据。我们认为有必要与大家分享我们熟知的两方面原因。

第一个是："大数据是作为一项研究措施开始的。在我们使用一个小的数据集并在一个受信任的环境中操作时，治理的概念并不是必须的。企业批准项目后，我们开始进行这项工作，我们会意识到这同任何其他企业IT项目没什么差别——需要对其进行治理。"这只是一个例子，不要对新数据仍沿用旧数据的思路，试图没完没了的改造治理计划，结果却导致比实际所需的更昂贵也更复杂。

我们得到的第二个最常见的回答是："我们在大数据项目中建立了很多新的功能、分析、模型等，所以，我们只是觉得在需要时就可以建立一个治理计划。在这里写一些代码加强安全性，在那里写一些代码来匹配数据，最后，悄悄地治理变成不是一个战术问题，而是想用一次性的方式解决这个问题，其代价是非常昂贵的。"

您开始从中看到一种现象了吗？这些组织不是从一开始计划他们的大数据系统治理工作，结果在后面时，他们必须面对大量的额外费用，以改造治理工作并将其纳入生态系统。

但是，那些充分利用其现有的信息集成流程和治理技术的企业，将会从新的大数据分析技术中，以一种安全的、可信的且紧跟潮流的方式获得最大的收益。

信息和分析方面的竞争

许多组织追求大数据分析，以寻找有突破意义的洞察，让他们在竞争中处于优势地位。他们在进行分析竞赛。但真的只能针对分析进行竞争吗？究竟分析的内容是什么？

现实的情况是，组织在信息和分析方面进行竞争，并且为了根据信息采取行动，所分析的数据必须是能够信任的。虽然我们都希望可以信任数据，但有些类型的数据，在这一点上未必是完全可能的。例如，您会相信一个随机的微信吗？这取决于具体情况，想想我们在第1章谈到的墨西哥总统选举的例子。从本质上讲，企业获取和创建他们信任的数据，但毫无疑问，他们肯定有一些数据是不能完全相信的。对于这种数据，有必要进行风险估计和评估。这是关键的一点，但常常被人们忽视。这种风险估计必须和与之相关的数据和决策制定过程联系在一起。事实上，根据这类数据上采取行动之前，必须让流程了解评估结果。集合和治理数据工作的最终目标是获取能够信任的信息。也就是

说，各种研究表明，在大多数组织中业务用户不信任他们在日常工作中使用的信息。在大数据世界中，更多的数据来源削弱了我们对数据的信任(如图10-1所示)。

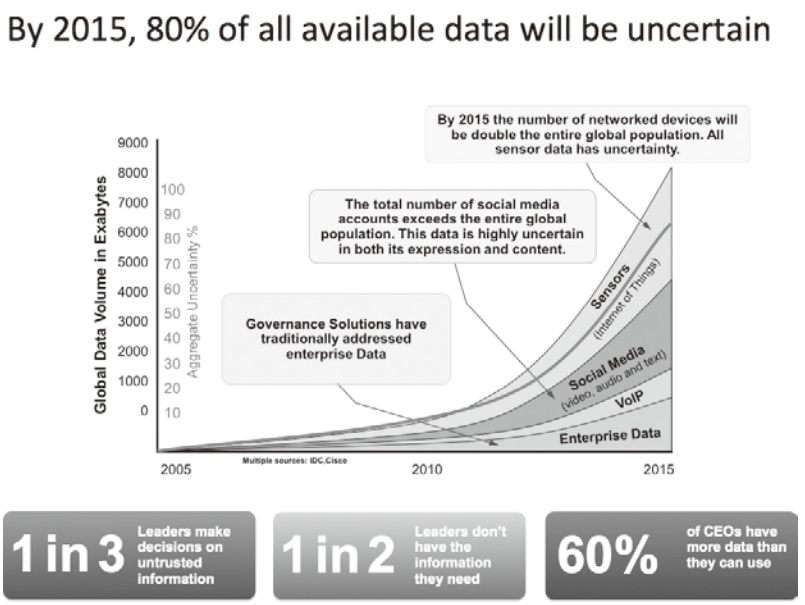


图10-1 对数据的信任度在下降，因为数据的来源和数量越来越多，让数据的不确定性越来越高。

By 2015, 80% of all available data will be uncertain: 到 2015 年，80% 的所有可用数据都将是不确定的数据

By 2015 the number of networked devices will……: 到 2015 年，连网设备将是全球总人口的两倍。所有传感器数据都具有不确定性。

The total number of social media accounts exceeds……: 社交媒体帐户的总数超过了全球总人口。该数据在表达和内容方面都具有极高的不确定性。

Governance Solutions have ……: 治理解决方案传统上可解决企业数据问题。

Global Data Volume in Exabytes: 全球数据量(EB)

Aggregate Uncertainly: 不确定性总计

Sensors: 传感器

Internet of Things: 物联网

Social media(video, audio and text): 社交媒体(视频、音频和文本)

Enterprise Data: 企业数据

Multiple sources: 多种来源

1 in 3 Leaders make decisions……: 三分之一的领导者根据不信任的信息制定决策

1 in 2 Leaders don’t have the ……: 二分之一的领导者没有他们所需的信息

60% of CEOs have more data……: 60% 的 CEO 拥有的数据要多于他们可使用的数据

开展分析所围绕的“净”数据源越多，带来的信任问题也越多。

可信数据的下降意味着，采用新应用程序和新技术来丰富分析生态系统时，会带来严重的障碍。信息集成和治理技术能够解决融会在大数据世界中的质量问题，方法是通过主动管理信息的治理，从而建立对信息的信任，并鼓励人们应用大数据分析。将信任和风险评估相结合，是行动之前必不可少的。

信息集成和治理的定义

市场已经积累了很多信息集成和治理定义。虽然介绍这些定义之间的细微差别不在本章的讨论范围之内，但我们提供了以下可行的定义，供本章剩余部分的讨论使用。

信息集成和治理(IIG)是一个组织如何处理信息的业务战略。它的核心定义了如何使用、共享信息和积极在组织内监控信息的策略。它涉及到策略定义、元数据管理、数据质量、信息集成、信息生命周期管理、隐私和安全性以及主数据管理技术。此外，还涉及到人员和流程，它们最终确定并执行各种治理策略。信息集成和治理的目的是建立和提供值得信赖的信息。

让我们来看看最后一个句子的本质：“治理的目的是建立和提供值得信赖的信息。”是什么让一个组织及其用户能够信任信息？我们相信这涉及到6个关键的因素。

- **信息是被理解的** 对信息的来源、价值和质量配置文件都有很好理解。
- **信息是正确的** 信息是标准化的，经过验证、确认和匹配的。
- **信息是全面的** 它不应该被分割，不应该有竞争版本的相同信息。
- **信息是最新的** 只存储最近和最相关的数

据，对旧的数据进行归档或删除。大数据经常提出生命周期管理方面的新挑战，大数据对时间往往非常敏感，而且会迅速失去其价值，这可能是因为对整个大数据组(如社交媒体数据)的生命周期治理不积极，没有对整个数据集进行分析，以及在分析完成后将其删除。这就是说，在大数据世界中也有这样的考虑，即存储语料库的数据来建立更好的预测模型。

- **信息是安全的** 对数据违规的保护水平(加密、新版本、安全性和监控)与公司治理要求相匹配。
- **信息是存档的** 必须跟踪各种信息的来源系统，以及应用到它的所有治理规则和转换，并且这些是可解释的，对最终用户可见。有时，人们将这个因素称为最终用户透明，因为每一个治理规则和流程都应当记录存档，并呈现给最终用户，以协助建立信任关系。

所有6个治理因素都需要应用到大数据吗？这取决于具体的使用情况：我们会谈论其中的两个，说明两种完全不同的治理方法。例如，分析大数据来检测欺诈模式，这必然要求记录数据的来源：也许可能涉及标准化和匹配来清洗重复的记录，了解全面的主客户数据，以匹配各种欺诈记录，通过屏蔽来保护敏感数据，甚至进行生命周期管理，从而用不同的时间间隔来归档每个记录或让记录退役。同时，使用大数据并通过社交媒体调查客户的情绪，需要的治理是完全不同的。这可能涉及到主客户数据与大数据的整合，以辨识客户，但社交媒体数据可能并不需要进行清洁、记录或对每个纪录进行生命周期管理——可能会在分析得出结论后就删除整个数据集。

信息治理流程

信息治理是一项业务战略，而不是一个IT项目——您买不到它，您必须对它投入极大的热情和精力，使它作为一个整体，成为组织中业务流程以及核心预期的一部分。因此，它需要改变组织的流程和人们对待信息的方式。为了实现可信任的信息这一目标，组织必须接受技术，也必须接受对流程以及人员的工作和激励方式的变更。只改变其中一个，是不可能实现这一目标的。更重要的是，这两方面的变更彼此要能够确保对方获得成功。一个全功能的治理平台让人们有可能修改流程，以主动治理信息，并且以人员为重点，这将鼓励他们接受和采用新技术。

在大多数情况下，针对传统关系型数据建立的信息治理流程同样适用于新的大数据源，所以我们想简单介绍一下信息治理的关键流程步骤，您可以将它用作大数据治理的路线图。

1. 识别要解决的业务问题。能够识别战略意义的业务需求，成为驱动技术革新的关键。如果首先构建技术，然后推动企业采用，这种方法永远不会奏效。虽然这一点

与大数据治理并没有直接的关系，但这是“任何”项目的必要措施，它自然地让治理成为确定项目范围的主题。

2. 提前获得有话语权的高管对任何大数据项目进行治理的支持，从而重申在整个组织中对大数据项目进行治理的重要性。该支持人通常与业务问题密切相关。由于治理真的是一个全企业范围的措施，所以您要评估支持人是否有政治影响力，可以将治理的重要性推进到整个企业中的其他项目。您还需要建立一个执行指导委员会，向其定期汇报治理大数据的进度。

3. 确定大数据流的哪些部分需要治理，以及治理到什么程度。治理规划的第一个关键步骤是，确定数据是否需要治理。要提出的问题包括：该数据是否为敏感的，它是否需要得到保护？这是通过法规进行治理的数据吗？该数据是否需要标准化？根据具体的用例，有些大数据可能只需要治理的某些方面，而有些数据可能完全不需要。但我们要强调，组织有意识地选择数据是否应接受治理，以及接受什么水平的治理，这是至关重要的。例如，可以将治理操作应用到所有数据(删除从上个月开始的所有Facebook数据)或单独的文件或记录(删除从上个月开始的无记名记录)。事实上，明确地评估一个大数据源并选择不治理它，这也是治理本身的一种形式。

4. 打造一个路线图。路线图涉及到需要治理的几个大数据项目。重要的是，至少要为治理措施确定实施项目的前三个阶段。为了保持从第一阶段到第二阶段的动力，组织必须在初期就规划出第二阶段。我们需要一直强调动力的重要性，它是至关重要的，在第一和第二阶段之间尤其关键，因为在这个时候可以决定企业措施是否能真正把握住机会，还是不了了之，变成只有一个阶段的遗憾。一个良好的第二阶段会充分利用第一阶段的某些方面，让您可以在第二阶段进展得更快。如果您在第一阶段使用了安全功能，也许在第二阶段会针对另一个大数据源充分利用它们；也有可能第二阶段以同一个大数据源作为重点，但增加了新的数据质量元素。关键是要保持简单性(无论是治理功能还是大数据源)，这样才能实现充分的利用。

5. 建立一个组织蓝图。蓝图是在组织不同层次的业务和IT参与者的一个矩阵式结构。工作团队包括确定治理策略的业务参与者，以及负责实施的IT参与者。它还包括一个执行指导委员会，行动小组向该委员会报告。高管支持者担任指导委员会主席。

6. 定义指标和目标。必须定义各种指标并设定目标，以跟踪成功情况。没有必要跟踪治理业务用例的每一个可能的指标；开始时只着重于十来个关键指标，它们应该是整体业务用例的成功和成就的准确预测指标。所跟踪和报告的指标应与业务价值或信息的可信性紧密相联。

7. 建立治理策略。行动团队确定项目的治理策略，并推动业务采用新的流程来治理数据。

8. 实施治理技术。虽然IT团队实施并集成治理技术来实现编辑、活动监控、权限使用等，但行动小组的业务代表可帮助推动人们采用可信任来源的数据。

9. 衡量指标和宣传。作为一个被忽略的新年决议，我们都知道这一步是非常重要的，但我们大多数人从来没有跟进。为什么呢？通常由于过于详尽的指标清单或在事后难以收集信息。选择容易收集的少数指标。宣传是至关重要的，因为那是创造成功以及因此而产生动力的时候。

信息治理是一个业务策略，推动对流程和执行这些流程的人员实现变更。大数据肯定会影响治理技术，但您所习惯的为关系型数据建立的治理策略和结构几乎完全可以转移到大数据时代。

IBM信息集成和治理技术平台

在IBM Information Management产品组合中的信息集成和治理(IIG)以大量业界领先的技术为代表，IBM已经将它们统一纳入一个可信的信息交付平台，其品牌名称为InfoSphere。IBM还投资了一个明确定义的信息治理流程，帮助组织规划自己的治理措施，以及它将如何影响人员和流程。虽然这些技术作为一个单一平台协同工作，但根据其用例而需要不同组件的组织，

可以利用特定的平台接入点。IBM IIG技术的主要功能如图10-2所示。

在本节剩下的部分中，我们会简单介绍这些核心平台功能、提供这些功能的相关IBM产品，以及它们在大数据世界中的应用。

IBM InfoSphere Business Information Exchange

InfoSphere Business Information Exchange (BIE)是IIG所有方面的一个核心组件，无论您需要提高数据质量、掌握数据、保护数据库，还是管理数据的生命周期，您都必须先理解数据并定义治理策略。BIE包含多个组件，可帮助组织了解和分析数据源，它包含多种独特的服务，可以确定信息治理项目的蓝图，

在它的数据源系统中发现数据，管理企业元数据，通过词汇表以业务友好的方式来表示该元数据，并定义治理策略等。

图10-2 IBM统一的信息集成和治理平台，可以帮助您将可信的信息交付给业务。

IBM InfoSphere Discovery

InfoSphere Discovery (Discovery)是一个BIE组件，为结构化数据源自动创建一个配置文件。Discovery可以确定源系统的逻辑模型(例如，通过查看物理数据和表来推断客户对象)。它也可以通过检查跨字段的数据关系推断出该数据源的一些变换或验证规则；它可以确定跨字段验证规则、允许值等。通过缩短在分析和记录现有数据源的劳动力密集型流程，这可以为集成流程提供很大的帮助。使用InfoSphere Data Explorer(前身为Vivisimo Velocity Platform)，IBM也能够发现、分析、索引、搜索和导航非结构化的大数据。

IBM InfoSphere Metadata Workbench

共享元数据是有效数据集成工作的基础。通过在企业级别共享术语和数据转换的公共定义，组织可以快速地将数据从一个系统翻译到另一个系统。元数据还记录数据的血统(数据来自哪里)、它要去哪里，以及此过程中它会发生什么事情。数据血统元数据是最强大的可信力量之一，因为将它暴露给业务用户时，它可迅速建立信任关系。InfoSphere Metadata Workbench拥有管理和共享结构化格式的企业元数据的功能。IBM InfoSphere Information Server还包含了分析非结构化的大数据源以及在所有大数据源上存储元数据的功能，还有更多即将推出的特性可以解决与大数据相关的一些新挑战。

IBM Business Glossary

元数据的有效共享依赖于业务和IT都同意的一个公共词汇表。这就是Business Glossary可以提供帮助的方式。它管理元数据的业务定义，将IT术语放进业务用户可以理解的公共术语中。此功能在大数据时代中甚至变得更重要。由于数据来自许多新的来源，它也带来了许多新的术语和格式。对公共词汇表的需求从来没有如此巨大。

信息治理中面临的最大挑战之一是制定各种策略，您确实必须在管理和治理数据的每个系统中制定它们。但是，如果您能在制定治理策略的同时定义元数据呢？Business Glossary可以让您做到这一点，在以通俗易懂的语言定义治理策略的同时定义元数据。Business Glossary提供业务驱动的治理，让业务分析师和用户能够定义各种治理策略。

IBM InfoSphere Information Analyzer

该组件分析来自源系统的数据的质量，从而分析系统本身。例如，计费系统是否为住址和电子邮件地址的一个可靠来源？这是建立信任的关键步骤。业务用户在试图建立信任时会提出的第一个问题是“此数据来自什么系统？”能够回答这个问题，并且能够收集有关系统是否真正可信赖的统计信息，都是建立信任的关键步骤。IBM InfoSphere Information Analyzer可分析源系统，并对这些系统中各种数据的质量进行分析。例如，记录的完整性、存储在字段中的数据是否真正符合标准的要求(例如，SSN应该是9位数字)，以及报告和可视化汇总的数据质量概况(准确和完整的记录百分比等)。可以在企业结构化的大数据源上利用此功能；在更大型的数据集中，可能要分析有代表性的数据集。信息分析的概念更适用于结构化数据，可以预期一个设定的结构，然后对其进行分析。在未来，还可以对半结构化数据源进行分析，并且利用文本分析可以研究关键字/值。信息分析的概念更适用于结构化数据。

IBM InfoSphere Blueprint Director

InfoSphere Blueprint Director (Blueprint Director)映射和管理活动的集成架构图。这是一个用于设计集成架构的可视化工具，它也让您可以深入到特定的集成步骤，并启动其他各种集成工具，例如，一个用于开发质量规则的用户界面。Blueprint Director也可用于映射出大数据集成架构，并且可以帮助您主动管理大数据集成架构。

IBM InfoSphere Information Server

有三种类型的数据集成：批量(或批处理移动)、实时以及联合。特定的项目(如大数据分析)往往需要综合所有这些类型，以满足不同的需求。

信息集成是大数据平台的一个关键要求，因为它让您能够充分利用来自现有投资的规模经济效应，并在扩展分析范式时发现新的规模经济效应。例如，试想在下一个最佳报价(NBO)应用中大量投资SQL的情况。如果此应用是基于SQL仓库的，通过调用一个Hadoop作业来查看与特征项清货条件有关的情绪趋势，就可以增强此应用程序。这有助于在执行此类优惠之前确定其接受程度。有能力利用熟悉的SQL调用一个函数，在Hadoop集群中产生一个MapReduce作业来执行这种情绪分析，这不仅是一个强大的概念；从提高投资效果的角度来看，这也是至关重要的。

也许情况是，您有一台机器在Hadoop集群上运行数据分析作业，并希望从管理折扣的系统中提取客户信息，希望能在特定日志事件和最终信贷之间找到较强的关联。这又回到了我们在第1章中所谈到的棒球比喻。大数据时代的特点是有各种专用引擎，并且它是这些引擎的协调工具(像棒球运动员用一只手投球，另一只手接球会更好)。这就是关键：信息集成让这一切可以实现。

信息往往是结构化或半结构化的，为了实现高吞吐量，就需要强大的处理引擎。您的大数据集成平台应该能均衡地优化不同的集成和转换需求，从 ETL(提取、转换和加载)，到ELT(利用目标系统处理转换，同时提供转换逻辑)，再到TELT(转换、提取、加载和转换)。

有些开发人员认为，新的技术(如Hadoop)可用于多种任务。从批处理和集成的角度来看，大数据世界的特点是大数据技术具有各种方法和学科。这导致了假定一切都可以围绕新技术构建的“构建心态”。如果您回想一下，当数据仓库行业还处于起步阶段时，许多IT专业人士试图建立内部集成功能。今天很少有人会这样做，因为已经有成熟的信息集成技术。同样的模式也适用于Hadoop，有些人认为它应该是集成或转换工作负载的唯一组件。

例如，有些人提出，它们应该只使用Hadoop来为数据仓库准备数据；这通常称

为ETL。但是，通用工具和专用工具之间有着巨大的差距，并且除了数据转换，集成还涉及到很多方面，如提取、发现、分析、元数据、数据质量和交付。组织不应该只利用Hadoop进行集成；而应利用成熟的数据集成技术来帮助加快其大数据部署工作。数据集成中将采用新技术，如Hadoop；例如，在ELT风格的集成(其中的T可以在数据仓库中由存储过程来执行)中，组织可以利用Hadoop进行转换处理。我们相信，您会发现需要将Hadoop引擎用作ETL/ELT战略的一部分，但您也将大大受益于专用转换引擎、大规模并行集成引擎的灵活性，它们可以支持多种转换和加载要求、集成到常见的运行时环境，以及InfoSphere Information Server (IIS)等产品所提供的公共设计调色板。事实上，该产品的并行处理引擎以及端到端的集成和质量功能比替代方法有着明显的总体拥有成本优势。

例如，如果转换完全是SQL操作，IIS可以把这些操作向下推送到IBM PureData System for Analytics设备(前身为Netezza)。您的集成平台应该能够根据需要自动生成在Hadoop基础架构或ETL并行引擎上运行的作业，并且能够以公共的作业定序器管理它们。IIS包括对Hadoop和大数据文件阶段(BDFS)容器的连接，以支持数据的持久性和检索。单一的数据集成平台(如由IIS提供的平台)可以为您同时提供功能和灵活性。IIS包含大量预构建的转换对象和数百个函数，都构建于并行执行环境之上，让您可以灵活地使用最适合手头任务的多种技术(包括Hadoop)。IIS集成HDFS作为数据交付的源和目标系统。IIS也可以模拟集成阶段内的特定集成任务，并指定要在Hadoop上执行的流程，这将充分利用Hadoop的MapReduce处理和低成本的基础架构。通常可将其用于ELT风格的集成，其中的T并不是由数据仓库存储过程执行，而是由一个Hadoop系统执行转换。IIS还集成了InfoSphere Streams (Streams)，它可以将Streams过滤的洞察或数据积累到一个临时数据文件中，然后将该文件加载到目标系统中(例如数据仓库，进行进一步的分析)。

高速集成到数据仓库将是关键，IIS也提供了这个功能。IIS大数据转换流的一个示例如图10-3所示。您可以看到，该作业分析高保真的电子邮件(存储在Hadoop中)，以了解客户情绪，然后将分析结果用于更新仓库(例如，Customer维度)；这是一个根据电子邮件分析进行风险分类的示例。

图10-3 利用大数据资产组合的一个数据流作业，其中包括在Hadoop中的源数据与DB2关系型数据的联接，以及各种转换，从而对风险进行分类

在今天的IT环境中，常见的其他集成技术(在大数据世界中仍是关键)包括实时复制和联合。实时复制、利用IBM InfoSphere Data Replication (Data Replication)等产品，涉及监测源系统并触发对目标系统的复制或变更。这通常用于低延迟的集成需求。Data Replication具有先进的功能，可以支持高速数据移动、冲突检测、系统监控，以及一个用于设计集成任务的图形化开发环境。此外，它与IBM PureData Systems系列集成，可实现高速的数据加载/同步，还集成了Information Server，以积累各种变更并批量地将数据移到目标系统。

数据联邦通过联合查询访问联邦存储中的数据。该方法常用于从多个系统中检索数据，或使用来自一个系统的信息补充存储在另一个系统中的数据。IBM InfoSphere Federation Server (Federation Server)访问和集成来自一组不同的结构化数据源的数据，不需要考虑它们驻留在哪里。它通过联接来自多个存储库的数据，以及通过InfoSphere Information Services Director公开信息即服务(IaaS)，从而实现混合数据仓库。从联合的角度来看，重要的是能够跨企业资产执行联合搜索(如我们在第7章介绍的Data Explorer技术)，以及利用SQL等查询API执行联合搜索。Federation Server可以集成Data Explorer，以便在一个整体大数据(结构化和非结构化数据)联合搜索和发现中提供结构化数据源的搜索和查询功能。

我们认为，组织不应该仅试图跟Hadoop提供企业集成；而应在合理的情况下充分利用成熟的数据集成技术，以帮助加快其大数据部署。通用工具和专用工具之间有一个巨大的差距，更何况集成还涉及到数据交付以外的其他许多方面，如发现、分析、元数据和数据质量。我们建议您考虑在大数据项目中使用IBM Information Server，以优化将大量结构化数据加载(通过批量加载或复制)到数据仓库的过程，并在需要时通过联合来扩展它；优化将结构化或半结构化数据加载到Hadoop；并优化信息的收集，这些信息经

过流分析的过滤和分析。然后，您可以将数据加载到关系型系统(如数据仓库)，跨关系型数据库执行联合查询，作为大数据联合和导航的一部分，并将数据源复制到Hadoop集群或其他数据仓库。

数据质量

数据质量组件可用于确保信息的洁净度和准确度。IBM InfoSphere Information Server for Data Quality (IIS for DQ)是一个市场领先的数据质量产品。它包含了创新的特性，如信息分析和质量分析、地址标准化和验证，并且它全面集成到InfoSphere Information Server平台中，支持质量规则开发、质量作业在Information Server的并行处理平台上的执行，以及与企业元数据组件共享元数据。数据质量的讨论通常包括以下服务:

- **解析** 分离数据并将它解析为结构化的格式。
- **标准化** 确定什么数据将放在哪个字段，并确保以标准格式(例如，一个9位数的邮政编码)存储它。
- **有效性** 确保数据是一致的；例如，一个电话号码中包含一个区号和表示其位置的正确数量数字。它可能还包括跨字段验证，如针对一个城市进行电话区号检查，确保它是有效的(例如，对于多伦多来说，区号416是有效的，415则不然)。
- **验证** 针对已验证的信息源检查数据，确保数据是有效的；例如，检查地址值的确是一个真实有效的地址。
- **匹配** 识别重复的记录，并正确地合并这些记录。

组织应确定是否需要在分析前对他们的大数据源进行质量检查，然后应用相应的数据质量组件。加载数据仓库时；加载和分析将被集成到数据仓库的新大数据源时；以及大数据分析依赖于更精确的视图(例如，反映客户洞察)时，即使在Hadoop中管理数据，大数据项目都有可能要求您专注于数据质量，以确保准确性和完整性。

主数据管理

主数据管理(MDM)为关键业务实体(如客户、患者、产品、零件、供应商、帐户和资产等)创建并维护单一版本的真相。MDM是记录的运营系统，它在大数据生态系统中起着重要的作用。IBM的InfoSphere Master Data Management是市场上最全面的MDM解决方案。

我们认为，MDM可以为大数据分析提供一个很有吸引力的起点，因为MDM的定义侧重于组织内最高价值的实体。许多组织想象，他们希望对社交媒体进行分析，以确定客户的情绪，但他们知道谁是自己的客户吗？他们知道谁是他们最好的客户？着手对一个精确主题进行大数据分析时，在大数据分析应用程序中利用MDM系统的知识往往

是合理的；例如，了解一组特定的可盈利客户的情绪和下一个最佳行动，而不是针对您的公司分析广泛的情绪。

MDM和大数据之间的集成点包括摄取和分析非结构化数据、创建主数据实体、将新的分析信息加载到MDM系统中、与大数据平台共享主数据记录或实体，作为大数据分析的基础，以及重用在大数据平台中的MDM匹配功能(如客户匹配)。您会发现，IBM是这一领域中的领导者，虽然我们不能评论在此问题领域中将会出现的集成点，但IBM计划继续根据真实的客户用例构建MDM和大数据之间的集成点。

如果分析的目标是精确而不是广度(汇总)——某个人客户或客户群、特定的一个产品或一组产品等，您应该考虑将MDM解决方案与大数据配合使用。此外，当大数据分析的输出应该是“运营化的”；具体而言，当所获得的洞察将作为运营系统中采取行动的依据时，请考虑将您的大数据集成到MDM系统中。例如，试想要挽留客户的下一步最佳行动应该是多渠道行动一致；行动标志(或洞察)应存储在一个MDM系统中。

数据生命周期管理

不要将大数据视为“新技术”(例如，Hadoop)。将它视为一种现象。而大数据现象的主要特征是这样一个事实，即您的每一个系统中的数据都在不断增长。未经检查的数据增长对您现有的系统(数据仓库、事务系统和应用程序)有着巨大的影响。数据增长可能会导致这些应用程序成本高昂并且性能较差。数据增长至“大数据水平”也影响了测试数据的管理。试想一下。每当您部署一个系统时，您都需要从生产环境生成测试数据系统，供开发部门进行测试等。通常情况下，数据是从生产环境复制的，并且随着总数据量的增长，测试数据环境同样也呈指数式增长。测试数据的第二个主要问题是确保安全和隐私——在非生产环境使用之前，屏蔽敏感数据。总之，现有系统中的数据增长在削弱它们，并且在“大数据”时代中，问题只会变得更糟，除非数据得到了积极的处理。

数据生命周期管理控制数据的增长，因此也控制数据的成本。它以两种主要方式管理数据生命周期。首先，它可以帮助进行数据增长管理，提供一个框架来分析和管理数据的生命周期，并以高度压缩和高效的方式主动归档数据。其次，数据生命周期管理对于适当的测试数据管理至关重要；具体是指创建大小合适的、接受治理的测试数据环境，以优化数据存储和成本。IBM InfoSphere Optim (Optim)产品家族是数据生命周期管理领域中的市场领导者。Optim包含了市场领先的数据增长管理功能，以及跨异构环境归档完整的业务对象，同时还可以通过查询方便地检索归档信息。InfoSphere Optim Test Data Management (Optim TDM)包含成熟的测试数据管理功能，可生成大小合适

的测试数据集，屏蔽功能可确保敏感数据得到保护，而自动化功能可支持用自助服务的方式生成测试数据集。

归档关系型数据仓库时，您已经在思考数据生命周期管理，确保只存储当前信息，从而随着数据量的增长而提高性能并降低成本。

为了确保符合数据保留和保护规定，并且能够审计对数据保留策略的符合性，生命周期管理也是必不可少的。在大数据世界中，低成本的引擎(如 Hadoop)提供了一个机会，用低成本的替代方案存储在线档案，以托管较冷的数据。虽然平台之间的可转换性正在提高，但生命周期管理和归档的业务流程是完全不同的挑战(换句话说，Hadoop可能是一个低成本的平台，但还需要多得多的功能才可以真正地管理数据增长)。这就是Optim的切入点。它管理生命周期和归档流程——发现和分析大数据，并跟踪生命周期里程碑(何时归档)、自动归档来自数据仓库和事务数据库的大数据、提供可视性和按需要检索和恢复数据的能力、确保归档数据不可修改，以防止数据错误，还有符合数据保留和可审计性的法律规定。Optim可以将归档数据存储在一个高度压缩的关系型数据库中，在文件系统上的一个归档文件中，并且该文件可以被加载到一个Hadoop文件系统上。Optim和Hadoop之间的后一个集成点是一个关键的集成点，通过将归档文件放到Hadoop系统上，它提供了低成本的存储，同时还允许对数据进行分析，以支持不同的用途，由此从归档文件得出洞察。难怪我们有些人将Hadoop称为新的磁带，这就是原因！

实现大数据项目时，测试数据管理应该是一个肯定要考虑的因素，以控制测试数据成本，并改善整体实施速度。Optim TDM为大数据系统(如数据仓库)自动生成并刷新测试数据，它与IBMPure Data System for Analytics的集成经过了优化。Optim生成合适大小的测试数据集，例如，一个100 TB的生产系统可能只需要1 TB进行用户验收测试。InfoSphere Optim还可以确保敏感数据在测试环境中被屏蔽。它会生成用于测试的真实数据(例如，将Jerome Smith at 123 Oak Street改为Kevin Brown at 231 Pine Avenue)，但它会保护真实的数据，以避免潜在的数据损失或误用。数据丢失的可能性在大数据的新时代中变得更加真实。更多的系统和更多的数据导致更多的测试环境和更大的数据丢失可能性。OPTIM TDM也有一个巨大的成本优势，利用在整个合适大小的环境中的自助服务数据生成，Optim TDM精简了测试数据流程。

大数据与数据增长有关，并且企业每一个系统中的数据都在增长，这是毫无疑问的。数据生命周期管理应该是最高优先事项，以帮助遏制未经检查的大数据增长，降低成本的数据，同时让应用程序更高效。

隐私性和安全性

在信息集成和治理讨论中有多个隐私和安全方面，其中大部分可以应用到大数据。无论敏感数据驻留在何处，您都需要保护它并阻止对其未经授权的访问。如果您必须为收集的某一类数据应用治理，不论该数据存储在文件系统(如HDFS)还是在关系型数据库管理系统(RDBMS)中，都有隐私和安全问题。例如，您的安全咒语(职责分离、关注点分离、最小特权原则，以及纵深防御)应用到存储在任何位置的数据。您将要考虑基于角色的安全性、多租户，通过反向代理(以及其他安全服务)减少表面区域配置，所有这些都是BigInsights可以提供给Hadoop的。

当然，值得注意的是，如果IBM InfoSphere Guardium (Guardium)是审计方面的业界领导者，通过其异构数据活动监测(DAM)服务，根据数据管理水平的活动发出警报，为什么不能对HDFS这样做呢？在2012年第三季度，IBM宣布Guardium开始支持Hadoop环境(NameNode、JobTracker和DataNodes)及其子系统项目(例如，Oozie)并提供DAM服务，让管理员能够清楚地了解谁做了什么、谁修改了什么等。在撰写本文时，目前有审计日志的大多数BigInsights组件都可以被监控，例如，HDFS命名空间操作、MapReduce(作业队列和作业操作、刷新配置)、HBase Region Server(数据库活动)、Hive和Avro。因为Guardium与开源组件配合工作，它能够与BigInsights和/或其它开源版本的Hadoop进行集成。Guardium也识别Thrift和MySQL协议(由Hive使用)。所有这些组件都可将现有的BigInsights审计日志发送给Guardium，并且可以利用它来满足合规性要求，如保存、处理、告警和报告审计日志。未来感兴趣的领域可能包括为HBase、BigInsights Web控制台等捕获审计日志。

数据库活动监测是有效治理大数据环境所需的一项关键功能。事实上，我们认为，对此功能的需求甚至超过了传统数据库环境中对治理的需求，因为目前Hadoop的治理控制通常较弱。

与传统数据一样，大数据可能在测试和生产环境都需要数据屏蔽。屏蔽数据是新的大数据技术中一个最大的问题，正如许多客户已意识到，他们可能会在测试和生产大数据环境中不小心暴露非常敏感的信息。您需要创建真实数据的仿真版本，但同时保护敏感数据值不会受到损害。对传递到HDFS或数据仓库的敏感数据进行屏蔽将成为(并且应该已经成为)大数据环境中一个紧迫的问题。IBM InfoSphere Optim Masking Solution通过在Hadoop系统中屏蔽数据来解决这一问题。事实上，Optim采用基于API的方法实现屏蔽，这意味着，任何系统都可以利用其先进的屏蔽功能，并在其处理中纳入屏蔽。好处是明显的——能够集中地定义屏蔽规则，并将其应用在多个大数据系统中。

模糊处理和用黑色涂掉文档内特定的敏感内容也将是关键；毕竟，如果您在HDFS中存储了包含敏感数据的电子邮件，可能需要修改此电子邮件。这里值得一提的是，在InfoSphere Guardium Data Redaction中使用了作为BigInsights和Streams的一部分的Text Analytics Toolkit。

小结：信任就是要将大数据转换成可信的信息

信息集成和治理是在任何大数据平台的设计阶段都应考虑的重要组件，因为从数量、速度、种类和真实性的角度来看，最终要大规模注入笨重的数据。客观而言，IBM是信息集成和治理领域中的领导者；正如您可能已经注意到，组成IBM InfoSphere Information Integration and Governance平台的一些产品得到了扩展，以支持IBM大数据平台；有一些目前已存在，有一些即将推出(虽然我们的律师希望我们将它留作惊喜，但我们相信，我们已经提供了一个水晶球，让您可以通过我们在本章详述的主题看到未来)。

虽然信息集成和治理创建了可信的信息，但我们真心希望您会记住，信息集成和治理不仅是一项技术；它还是一项业务战略和相关的循序渐进的变更流程，可满足企业范围的治理目标。在您设计自己的第一个大数据项目时，关键是要确定项目的哪些部分需要接受治理，以及接受哪种程度的治理。

治理流程的第一步也是最关键的一步，是决定大数据是否需要接受治理，以及它需要接受哪种程度的治理。这对于任何大数据项目都应该是强制性的任务。没有做出清醒的决策将可能导致一场灾难……如较低的采用率，或更糟的是，成为只有一个阶段的绝唱。这是信息集成和治理技术的真正价值，它可以确保您的大数据措施获得成功并推动人们采用它。大数据分析的成功取决于可信信息的供应。

11 在企业中集成大数据

有一件事是肯定的——大数据技术不应该是一个竖井。它必须在企业内部集成，以充分实现其价值。让我们仔细想想。组织需要一个技术平台来开发大数据分析应用程序。这些应用程序必须能够与这个大数据平台交互，以利用特定的大数据功能。这里是第一个集成点：在大数据平台上构建的分析应用程序。

大数据平台内包含许多功能。为了充分满足需求，通常需要多个大数据组件。这是第二个集成点：大数据平台组件和产品必须彼此集成。

毕竟，如果不打算根据从大数据收集到的洞察采取行动，大数据措施又有什么意义呢？

在我们的经验中，为了根据洞察采取行动，通常必须与可利用洞察的其他一些应用程序共享洞察。因此，有第三个集成点：企业应用程序和存储库。

平台应该是开放的，使更广泛的生态系统中的合作伙伴技术可以扩展解决方案。这是最后一个集成点：大数据平台的插件组件。因此，我们在本章中要探索的四种类型的集成分别是：分析应用程序、平台内的集成、与其他存储库的企业集成，以及平台插件。

很多客户在向我们咨询其现有技术投资和新的大数据项目的集成点时，已踏上了正确的轨道；的确，这听起来一直是令人鼓舞的，因为它意味着，客户的思路已经超越了极少产生价值的“科学实验”阶段。需要注意的是，还有更多集成点是我们无法在本章中——介绍的，我们鼓励您使用本章作为一个快速起点，而不是集成选项和模式的一个最终讨论。

分析应用程序集成

IBM大数据平台是从头开始设计的，以简化在它上面运行的应用程序的开发流程，并通过一组经过优化的底层大数据技术提供分析的优势。本节介绍IBM Business Analytics软件以及与IBM大数据平台集成的其他应用程序。由于大数据生态系统涵盖了200多个业务合作伙伴，还有许多我们没有介绍的应用程序，当然，这个列表每天都会变得更长。

IBM Cognos软件

IBM Cognos BI是IBM Cognos Enterprise的一个组件，它是一个被广泛部署的商业智能软件，通过与IBM大数据平台的集成而得到扩展，与大数据配合使用。在撰写本文时，Cognos可以访问存储在Hive的数据，并将数据用于报告(有更多特性即将推

出，敬请关注)。Cognos BI与InfoSphere BigInsights (BigInsights)、InfoSphere Streams (Streams)、DB2、IBM PureData System for Analytics(前身为Netezza)和IBM PureData System for Operational Analytics(前身为IBM Smart Analytics System)集成并获得认证。

IBM Cognos Real Time Monitoring (Cognos RTM)是IBM Cognos Enterprise的一个组件，该软件对来自Streams的实时流分析提供可视化和分析。可视化是大数据带给业务分析师的主要挑战之一；事实上，一些大学如今提供了一定程度的大数据可视化。在本书中，我们一直强调运动数据是让IBM大数据平台与众不同的因素，因为该领域的其他厂商在解决速度(四个大数据特征之一)问题时并没有投入与IBM对此问题领域相同程度的关注。如果速度是大数据的一个非常重要的方面，显而易见的是，需要将运动数据集成到商业智能，让用户可以获得对过去、现在和未来的洞察。

IBM Cognos Consumer Insights (CCI)是一个富社交媒体分析应用程序。它允许用户分析、可视化和报告消费者情感以及来自社交媒体不断变化的市场主题。IBM CCI集成了IBM的大数据平台，尤其是BigInsights。BigInsights为CCI提供了可以处理大量原始社交媒体数据的功能和灵活性。

IBM Content Analytics with Enterprise Search

IBM Content Analytics(ICA)为分析非结构化的内容提供了领先的功能。但是，如果数据集增长得太快，无法处理(就像我们知道的，它们经常会这样)，会发生什么呢？BigInsights与ICA集成，并且两者配合工作，以确保飞速上升的数据量不会成为一个问题。IBM Content Analytics还集成了基于分析的IBM PureData Systems，这样它们就能够共享洞察和信息。利用IBM InfoSphere Data Explorer(通过Vivisimo收购而获得的技术，我们在第7章对它进行了介绍)，ICA提供了大量非常强大的非结构化信息处理、发现和消费选项。

SPSS

IBM SPSS软件是一个广泛部署且成熟的预测分析产品组合。客户希望利用SPSS与IBM大数据平台进行交互，这并不奇怪。静止建模是很好理解的，但如果您想在极端的情况下利用预测分析——无论是在销售点、对冲基金交易员的Bloomberg终端，还是急诊室监控设备，都在数据形成时进行预测，那会怎么样呢？毕竟，预测建模的意义在于改善成果。SPSS和Streams的组合就是答案，因为它让您能够对静止数据构建预测模型，之后利用运动数据对其进行评分。为了支持这一强大的使用模式，可将SPSS预测模

型直接导出到Streams运行时环境。此外，Streams支持使用Predictive Model Markup Language (PMML)定义的模型，SPSS也支持此类模型。Streams提供了在庞大的规模和范围内实时运行两种预测模型的能力。几乎整个IBM大数据平台都集成了PMML。例如，您可以在SPSS构建模型，然后在IBM PureData System for Analytics设备中运行它们，以非常快的性能执行深度分析。

SAS

SAS业务分析应用程序与IBM大数据平台集成。具体来说，SAS与IBM PureData System for Analytics集成。这种集成让SAS评分和建模可以在此PureData System内运行，这对于将分析带给数据而言会更高效，当然它也受益于基于分析的PureData System为在大型数据集的深度分析查询所提供的巨大优势。

Unica

Unica是进行跨渠道营销活动管理和营销绩效优化的一个先进解决方案。它完全集成了IBM PureData System for Operational Analytics和IBM PureData System for Analytics，并且可用于在这些平台上执行营销活动分析和报告。IBM PureData System带来的好处是很明显的——能够对大数据量执行深度分析，并以闪电的速度得到您所需的答案，或支持运营分析工作负载。Unica解决方案利用大数据平台的力量，可以更迅速地分析更多的数据，以产生更多有针对性的营销活动。

Q1 Labs: Security Solutions

安全性和大数据的主题已达到白热化。最紧迫的问题是，“如何利用大数据技术扩展现有安全解决方案？”2011年，IBM收购了Q1 Labs及其QRadar Security Intelligence Platform (QRadar)，该产品提供了一个统一的架构，用于收集、存储、分析和查询与日志、威胁、漏洞和风险相关的数据。为了响应客户的需求，IBM的安全性和大数据团队一直致力于集成点，从而在两个平台之间传递洞察和分析，带来向外扩展的选项和更大范围的态势感知。QRadar作为一个独立的技术非常引人注目，但与大数据结合时，我们正在实现以前从不可能的事情，帮助锁定我们在企业客户领域地位。

IBM i2 Intelligence Analysis Platform

IBM i2提供了一个可扩展的、面向服务的环境，其目的是与您现有的企业基础架构集成。该平台可推动和支持运营分析，提高态势感知，并在整个组织内提供更快、更明智的决策。自收购以来，IBM一直忙于集成i2平台与IBM大数据平台，并扩展i2最终用户可以发现和探索的数据。i2高度可视化和直观的最终用户分析可以帮助您识别和探

索各种模式，然后在快速发展的环境中采取行动。

Platform Symphony MapReduce

2011年，IBM收购了Platform Computing，包括他们的Platform Symphony MapReduce技术：一个受Hadoop启发的运行时，专为跨多个同时短暂运行的作业需要低延时或复杂的调度逻辑的应用程序而设计。BigInsights是为Hadoop提供的领先IBM产品，并且Platform Symphony被认为是该产品的一个相辅相成的组件。

有趣的是，这些产品的集成工作已经完成。例如，BigInsights可以利用Platform Symphony MapReduce作为在它自己的集群上运行的低延迟分布式运行时组件。这让拥有特定的低延时或繁重的计算工作负载的客户可以利用该技术作为BigInsights部署的一部分。可以将BigInsights部署在Platform Computing网格内的一个或多个实例上，从而利用其支持异构应用程序的技术计算能力。现有的Platform Symphony客户(有很多)可以将BigInsights 添加到他们现有的网格中，并在一个公共的资源池中管理多种类型的工作负载。这消除了为BigInsights建立一个单独网格环境的需要，并支持跨计算密集型和数据密集型应用程序进行资源共享。

IBM大数据平台内的组件集成

IBM大数据平台的所有组件都已实现全面的集成。在本节中，我们简要地强调部分此集成，并告诉您在哪些章节可以找到更多的详细信息。

InfoSphere BigInsights

BigInsights在整个IBM大数据平台中实现了深入的集成。BigInsights可以接收来自Streams的实时输入，如由部署到Streams运行时的分析算法所确定的洞察和决策，然后再对更深入的数据集执行更多分析(例如，测试和改造算法，以提升回Streams)

BigInsights还集成了基于分析的IBM PureData Systems和其他IBM数据仓库解决方案，如InfoSphere Warehouse。这让BigInsights可以将数据写入到一个仓库，也可以从仓库接收输入。使用SQL也可以访问在BigInsights中的数据。提供SQL包装的用户自定义函数的数据库内函数让您能够从数据库内调用一个BigInsights作业。(但是请注意，物理定律在这里是适用的：您不会打算利用这些函数通过网络将数万亿行数据下载到您的仓库中，您更有可能希望调用一个作业并检索一个已评过分的结果)。使用BigInsights分析原始数据或执行实验分析，然后与数据仓库分享洞察时，此集成点是有用的。

BigInsights和InfoSphere Information Server (Information Server)之间也存在集成。Information Server可以向Hadoop提供信息或从Hadoop读取数据，并将它集成到任何数量的其他系统。BigInsights还集成了Guardium，以提供实时的系统监控和安全性。InfoSphere Optim Masking Solution可以屏蔽BigInsights数据，确保它受到保护并且是安全的。InfoSphere Optim可以在BigInsights中存储已归档的数据，让该数据可用于分析。IBM Research已经开发了MDM和BigInsights之间的集成点，包括能够分析原始数据，并确定哪些主实体被加载到MDM。您可以在第5章到第8章中找到关于此集成的更多详细信息。

InfoSphere Streams

Streams可以向BigInsights提供实时更新，这可能是Streams操作的结果或已经过评分的SPSS模型。BigInsights往往与Streams结合使用，以提供更深入的模型分析，之后在Streams中可以对这些模型进行提升评分(在第2章中讨论了一个类似的使用模式)。

Streams通过一组数据库适配器集成多个数据仓库解决方案，这些适配器让Streams能够将洞察写到目标仓库系统，或在目标仓库系统中筛选数据。我们经常会看到Streams集成一个仓库，以实现对结构化数据进行更深入的分析、建模和发现。事实上，Streams设置了经过负载优化的高速连接器，它们与基于分析的IBM PureData Systems实现了深入的集成。第6章讨论了这些细节。

最后，Streams可以同Information Server共享数据，或向其提供输入，它也可以通过Information Server按需要集成任意数量的系统。

数据仓库解决方案

数据仓库解决方案充分集成了IBM大数据平台的信息集成和治理(Integration and Governance，IIG)方面：信息集成实现数据的高速加载和卸载；数据质量用于标准化和完善数据；数据生命周期管理实现数据归档并控制数据增长，并且实现大小合适的测试数据环境；MDM系统用于提供业务实体的单一视图；还有隐私和安全监视，以屏蔽数据和监控数据库活动。IBM PureData System for Analytics和IBM PureData System for Operational Analytics已优化了对许多IIG解决方案的集成点(我们将在本章的后面介绍这些集成点)。当然，如前所述，IBM大数据平台的数据仓库解决方案也拥有与BigInsights和Streams的集成点。

Advanced Text Analytics Toolkit

IBM将Advanced Text Analytics Toolkit作为其大数据平台的一部分。此工具包中包括各种示例、可立即在生产中使用的内置提取器、集成开发环境、名为Annotated Query

Language (AQL)的文本提取声明性语言、文本分析引擎和相关的基于成本的优化器等。Advanced Text Analytics Toolkit随BigInsights和Streams提供，让您能够对静止数据开发文本分析模型，然后透明地将它们部署到运动数据。在第8章中可以找到有关此工具包的更多详细信息。

InfoSphere Data Explorer

IBM InfoSphere Data Explorer (Data Explorer)是IBM在2012年年初收购Vivisimo得到的产品集的新名称。Data Explorer提供对整个企业中大数据资产的安全和高度准确的联合搜索与导航。它可以在两种情况下与BigInsights和Streams集成: 利用Data Explorer及其可视化界面可以访问和共享来自BigInsights或Streams的数据，Data Explorer也可以直接从BigInsights使用数据。Data Explorer也可以与InfoSphere Federation Server集成，并成为其结构化数据存储库的联合索引的一部分。在第7章中可以找到有关此大数据技术的更多详细信息。

InfoSphere Information Server

IBM InfoSphere Information Server (Information Server)是IBM大数据平台的一个组件。Information Server通过它与多个企业系统集成的能力提供了一个企业网关。BigInsights与Information Server实现了集成，可以将数据加载到BigInsights，或将数据从BigInsights移动到其他企业系统。Streams也可与Information Server集成，让组织能够获得实时洞察，并以指定的时间间隔将该数据加载到目标系统。

Information Server与IBM PureData System for Analytics实现了深入的集成；它拥有将基于SQL的转换推送到其现场可编程门阵列(FPGA)协助的处理架构(有关详细信息，请参阅第4章)的“智慧”。Information Server还与IBM PureData System for Operational Analytics相集成。许多客户被IBM PureData System吸引是因为其方法的纯粹简单性。将Information Server集成到IBM大数据平台中的工作确保客户能够尽快实现可信信息基础架构的好处。在第10章中，我们提供了一个示例来说明Information Server的设计画布，用拖放手势来创建一个包括Hadoop数据源的流。

InfoSphere Data Replication与基于分析的IBM PureData Systems集成，通过其低影响、低延时的数据库监控，向两个存储库提供实时的复制数据。IBM的Information Integration(信息集成)可以满足大数据项目的批处理和实时集成的要求，因为它可捕获丰富的设计和运营元数据，以支持数据血统和数据治理分析。这让业务和IT用户可以了解企业数据如何在大数据平台内使用。因为大数据平台可用这种方式利用信息集成，所

以大数据项目可以充分利用几乎任何重要数据源。

InfoSphere Master Data Management

IBM已经证明BigInsights与IBM InfoSphere Master Data Management (MDM)产品套件可以配合使用。这项工作是由IBM Research团队与一些客户合作完成的，这些客户希望依靠来自大数据源的事件和实体解析，填充企业中的主配置文件。BigInsights用于分析原始数据，并为实体分析算法提取实体(如客户和供应商)。然后，该数据被进一步完善(例如，识别实体之间的关系)，之后被加载到MDM系统。

MDM概率匹配引擎也同BigInsights进行了集成，从而实现对大数据集的匹配。原始数据被细化和结构化，然后与其他记录进行匹配。许多客户对大数据与MDM的配合使用提出了疑问。其中的关联是很自然的。将MDM视为流程的“挡书板”；它可以向大数据系统提供清洁的主数据实体，从而进一步的分析，而从大数据系统收集的洞察可以反馈回MDM，帮助用户采取行动。

InfoSphere Guardium

InfoSphere Guardium Database Security (Guardium)是IBM领先的数据活动监测(DAM)解决方案，其好处最近被扩展到Hadoop。Guardium与BigInsights以及开源Hadoop的集成使其可以监视Hadoop系统，并确保企业内大数据的安全。

当然，Guardium很长一段时间以来一直与关系型系统集成，使其能够监测和保护结构化的大数据。正如我们之前所述：如果一个重要的数据仓库环境需要治理，在大数据的范畴中也同样需要它。我们在第10章中介绍Guardium如何用于保护Hadoop的一些细节，以及更多详细内容。

InfoSphere Optim

InfoSphere Optim (Optim)产品套件提供业界领先的数据增长管理、测试数据管理和安全功能。长期以来，它一直通过归档帮助数据仓库系统控制大数据增长。它还通过生成合适大小的测试数据集来确保实现高效的测试数据管理。此外，它通过屏蔽、在测试环境中使用逼真的(但不是真实的)数据，确保了敏感的测试数据的安全性。

所有这些功能都可以扩展到新的大数据技术。例如，Optim与Hadoop配合：它可以屏蔽在Hadoop系统内的数据。大数据系统也有敏感数据，有时甚至比其他系统更多。屏蔽可以确保敏感的大数据不会受到损害。Optim还可以从其他目标系统(运营应用程序、数据仓库等)提取已归档的数据，并将该数据存储在Hadoop中，这是可以对它进行分析的地方。

Optim还深入集成了多种网络的基于模式构

建的IBM PureData Systems，为这些系统提供经过优化的连接器，以归档数据，并生成合适大小的测试数据部署。我们在第10章中讨论了许多这些集成点。

WebSphere Front Office

IBM WebSphere Front Office (Front Office)组成的集成组件用于接收、整合市场数据，并将市场数据分发到金融机构内的前、中、后台应用程序。这一功能强大的产品在几乎所有客户环境中都提供了灵活性和可靠性。Streams与Front Office的集成为Streams提供了实时洞察和数据提要。

WebSphere Decision Server: iLog Rules

大数据处理往往需要复杂的规则来对数据进行分析，并且在过去的几年里，规则部署理所当然地获得了大量的企业关注。您的大数据环境可以利用WebSphere Decision Server (iLog Rules)来构建更智能的应用程序流。我们的研究团队已经展示了如何使用iLog Rules与Streams并根据iLog Rules所传递的条件和矢量来分析数据流(在本书出版后，我们将会就这一领域提供更多信息——我们的律师目前尚不允许我们详细介绍它)。

Rational

在大数据环境中，作业和应用程序开发生命周期管理往往被忽视，这是一个代价高昂的错误。Eclipse是一个可插入的集成开发环境(IDE)，它是由IBM发明并捐献给开源社区的。Eclipse在开源的大数据技术以及IBM大数据平台中得到了广泛使用。

IBM大数据平台为流处理语言(SPL)流、SQL、Java MapReduce、Hive、Pig、Jaql、基于AQL的文本分析等的开发提供了插件。可以从开源的Eclipse安装或从Rational Application Developer (RAD)等基于Eclipse的增值产品安装这些插件。IBM大数据平台利用针对Streams和BigInsights的开箱即用插件来扩展Eclipse IDE，这些插件可以充分利用底层的Eclipse项目工件。反过来，又可以通过版本控制(如Rational ClearCase或Rational ClearQuest)、协作(如Rational JAZZ)或企业级生命周期管理等扩展的Rational工具集来管理这些工件。

数据存储库级别的集成

如果要让大数据平台有效地工作，必须在平台内的各种存储库之间共享数据。本章和整本书中都介绍了多个集成点(参见第5章和第6章)。记住，业务流程并不是在大数据平台中启动和停止的——必须与其他存储库和平台共享数据。

企业平台插件

平台必须是一个开放的系统，第三方产品才能插入其中。IBM的大数据平台正是如此。

在撰写本文的时候，IBM拥有超过100个大数据合作伙伴！我们在本节的剩余部分并不打算详述细节，而是介绍一些集成合作伙伴类别。

开发工具

IBM的多个合作伙伴拥有针对特定流程和组件的开发工具。有些合作伙伴为BigInsights提供专用开发工具。IBM拥有提供大数据可视化功能的合作伙伴，如Datameer；还有针对基于Hadoop的数据分析为分析师提供开发和可视化工具的合作伙伴，如Karmasphere。

分析

正如本章前面所述，IBM内有几个分析应用程序和引擎被集成到大数据平台。也有一些分析引擎可以作为组件被集成到大数据平台。为了支持客户需求，有几个开源分析组件已经被集成到大数据平台：视频分析、音频分析和统计分析等。

可视化

有多个IBM合作伙伴专注于大数据的可视化。其中有些专注于Hadoop系统数据的可视化，另一些则专注于数据仓库应用中结构化数据的可视化，还有一些专注于运动数据的可视化。在上一节中，我们已经介绍了一些分析应用的示例，如Cognos Consumer Insights (CCI)，它能够可视化社交媒体数据，以了解消费者情绪和趋势分析。事实上，某著名的食品和饮料公司最终检测出了针对其最新环保包装变化的负面情绪，该包装遭遇了人们用数千个YouTube视频和Facebook页面表现出来的负面情绪，消费者通过这些渠道抱怨新包装非常吵，并且该Facebook页面有超过50,000个用户关注！他们使用CCI发现用户说了些什么。CCI在BigInsights之上运行，用于发现和分析为什么人们会说这些话。

小结

大数据平台及构建其上的应用并不是为了成为竖井。而是要成为可操作的，从大数据获得的洞察必须是共享的，并且集成能力最强的平台就是可以提供最大价值的平台。IBM的大数据平台拥有与分析应用程序、企业应用程序、企业存储库，以及各种平台插件等许多集成点，使其在功能和集成能力方面都成为最全面的大数据平台。

图11-1 说明了集成的效果，并展示了IBM在其大数据平台内已经完成的工作。

在本例中，正在分析的是多个客户交互数据源。在BigInsights中对Web站点日志和分析社交媒体提要进行分析。

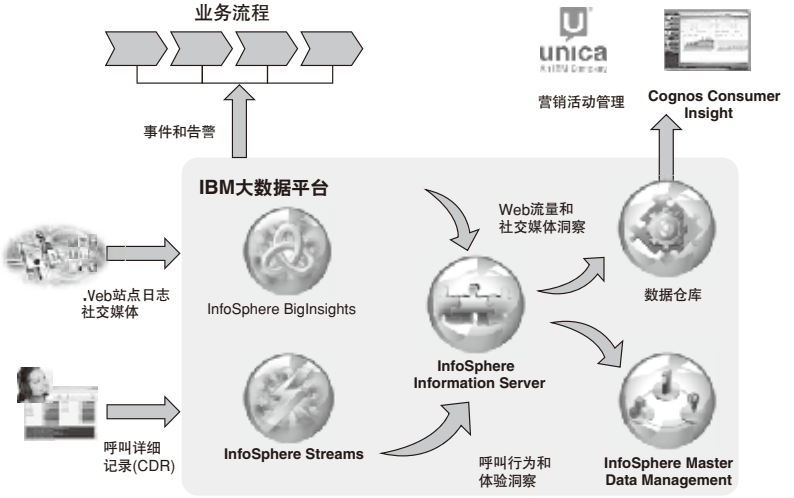


图11-1 在一个全方位的多渠道客户情绪分析项目中的集成价值

业务用户可以使用Cognos Consumer Insights来可视化该分析。Information Server将所产生的任何洞察馈送到数据仓库，从而可进行进一步的建模和分析工作。同时，Streams对实时的呼叫详细记录(CDR)进行分析，Information Server将所产生的任何洞察馈送到数据仓库。MDM技术将客户的单一视图作为BigInsights、Streams和数据仓库的分析起点，它也可以是可操作的任何洞察发挥作用的地方。最后，营销管理系统(如Unica)与数据仓库解决方案集成，以执行深入的分析(例如，客户挽留活动优惠)以及执行各种营销活动。

大数据技能的其他资源

BigInsights Wiki

依靠大量IBM专家、计划和服务，可以帮助您将自己的大数据技能提高到一个全新的水平。通过BigInsights wiki参与我们的在线社区。查找BigInsights白皮书、视频、演示、下载、社交媒体站点的链接、所有最新消息等。

请访问**ibm.com**/developerworks/wiki/biginsights

信息管理书店

查找本书的电子版、市场上最翔实的信息管理图书的链接，以及提供有价值的链接和优惠，以节省您的资金并提高您的技能。

请访问**ibm.com**/software/data/education/bookstore

大数据大学

按照您的节奏，在您的地方了解Hadoop和其他技术。大数据大学提供有用的在线课程、教学视频和练习，帮助您掌握全新的概念。课程完成的标志是最后的考试和证书。

请访问**bigdatauniversity.com**

IBM Data Management杂志

IBM Data Management杂志提供了各种全新交互的、高度动态的内容，如网络研讨会和视频。它还提供了一个平台，在其上创建一个由世界领先的信息管理专业人士组成的强大社区，您可从中了解来自整个行业内更广泛的意见。

请访问**ibmdatamag.com**

IBM Certification and Mastery Exams

查找业界领先的专业认证和Mastery考试。针对BigInsights (M97)和InfoSphere Streams (N08)提供了Mastery考试。

请访问**ibm.com**/certify/mastery_tests

IBM培训服务

查找经济高效且绿色的在线学习选项，如专属现场培训，以及传统的教室教训，均由我们经验丰富的世界级讲师授课。

- InfoSphere BigInsights Analytics业务分析师培训(DW640，2天)
- InfoSphere BigInsights Analytics程序员培训(DW651，2天)
- InfoSphere BigInsights基础(DW611，2天)
- InfoSphere Streams v2的管理(DW730，2天)
- InfoSphereStreams 编程(DW722，4天)

有关大数据培训服务的概述，请访问

ibm.com/software/data/education/bigdata.html

我们为了支持大数据所提供的软件服务的大量信息，请访问

ibm.com/software/data/services/offerings.html

Information Management订阅和支持

访问屡获殊荣的IBM支持门户，查找Information Management产品的技术支持信息，包括下载、通知、技术文档、Flash、警告等。

请访问**ibm.com**/support