IBM软件

Information Management

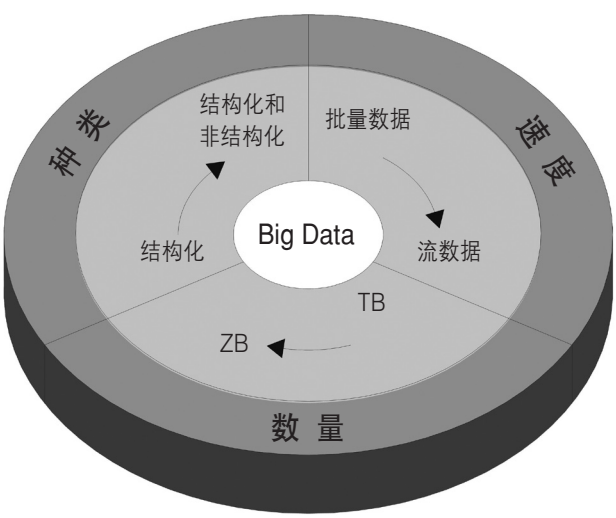
理解大数据

一项IBM调查发现，如今有超过一半的业务领导认识到他们无法获取完成自己的工作所需的洞察。尽管如今公司有能力存储任何信息并且正在以前所未有的方式生成信息，但这两方面相结合，带来了一个真正的信息挑战。这是一个复杂的谜题：如今的业务人员能获取比以往更多的潜在洞察，但由于这个潜在的数据金矿堆积成山，企业可处理的数据比例正在迅速下降。

1. 什么是大数据

可用3个特征来定义大数据：数量、种类和速度(如图所示)。这些特征相结合，定义了我们在IBM所称的“大数据”。他们创造了一种需求，那就是使用一类新功能来改善当今的做事方式，提供对我们现有的知识领域和驾驭其能力的更有效控制。

如今存储的数据数量正在急剧增长。在2000年，全球存储了800,000 PB的数据。当然，如今创建的大量数据都完全未经分析，这是我们尝试使用BigInsights解决的另一个问题。我们预计到2020年，这一数字会达到35 ZB。单单Twitter每天就会生成超过7 TB的数据，Facebook为10TB，一些企业在一年中每一天的每一小时就会产生数TB的数据。



图：IBM按数量、速度和种类或者就是简单的V 3来定义大数据

现在经常听到一些企业使用存储集群来保存数PB的数据。这里我们将列举一些可能的事实：在您阅读本文字时这些数据估计已过期，在您阅读完后将您的数据增长速率知识告诉朋友和家人时，这些数据会进一步过期。

停下来想想，毫无疑问我们正深陷在数据之中。如果我们可跟踪和记录某个事物，我们通常会这么做。(注意，我们没有提及分析已存储的这些数据，这将是一个大数据主题——对于我们跟踪但未用于决策制定的数据，这是新发现的用途。)我们存储所有事物：环境数据、财务数据、医疗数据、监控数据等。例如，从手机套中拿出您的智能电话会生成一个事件；当您的市郊火车到站开门时，这是一个事件；检票登机、打卡上班、在iTunes上购买歌曲、更换电视频道、使用电子收费公路——每一项操作都会生成数据。还需要更多数据？明尼阿波利斯的圣安东尼瀑布大桥(在2007年垮塌后被I35W密西西比河大桥取代)在重要位置布置了200多个嵌入式传感器来提供一个周密的监视系统，它会收集所有类型的详细数据，甚至温度变化和大桥对这一变化的具体反应都可供分析。您一定发现了其中的重点：现在的数据比以往更多，仅仅从个人家庭电脑的TB级存储容量即可看出。就在10年前，我们知道的超过1 TB的数据仓库屈指可数，这足以表明数据量发生了变化。

多样性是生命的调味料

与大数据现象有关的数据量为尝试处理它的数据中心带来了新的挑战：它的种类。随着传感器、智能设备以及社交协作技术的激增，企业中的数据也变得更加复杂，因为它不仅包含传统的关系型数据，还包含来自网页、Web日志文件(包括单击流数据)、搜索索引、社交媒体论坛、电子邮件、文档、主动和被动系统的传感器数据等原始、半结构化和非结构化数据。而且，传统系统可能很难存储和执行必要的分析，以理解这些日志的内容，因为所生成的许多信息并不适合传统的数据库技术。在我们的经验中，尽管一些公司正在朝大数据方向大力发展，但总体而言，大部分公司只是刚开始理解大数据的机会(以及如果不考虑它会有什么风险)。

当我们回头看看我们的数据库生涯时，有时会羞愧地发现，我们将大部分时间都花在仅20%的数据上：格式整齐且符合我们严格模式的关系类型。但事实是，全球80%的数据(越来越多的这类数据创造了新的种类和数量的记录)是非结构化的，或者至多是半结构化的。如果查看Twitter源，您会在其JSON格式中看到结构——但实际的文本不是结构化的，而且理解这些内容会得到回报。视频和图片不能轻松或高效地存储在关系型数据库中，某些事件信息可能动态地更改(如天气模式)，它们不太适合严格的模式。要利用大数据机会，企业必须能够分析所有类型的数据，包括关系和非关系数据：文本、传感器数据、音频、视频、事务等。

多快才算快？数据的速度

有效处理大数据需要您在数据变化的过程中对它的数量和种类执行分析，而不只是在它静止后执行分析。考虑从跟踪新生儿健康状况到金融市场的各种示例；在每种情形下，他们都需要以新的方式处理不同数量和速度的数据。大数据的速度特征是让 IBM 成为您最佳大数据平台的一个重要因素。我们将它定义为一种从单纯的批量洞察(Hadoop风格)到与动态传输的洞察相结合的批量洞察的内含式转变，IBM可能是唯一未将速度局限于数据生成速率(它实际上是数据数量特征的一部分)的供应商。

现在想象这样一种结合式的大数据平台，它可利用两个领域的优点，实时传输洞察，以获得基于新出现数据的进一步研究结果。正如您所想的，我们相信您会跟我们一样，对IBM大数据平台所提供的独特主张激动不已。

大数据平台允许您将所有数据存储为其原生的业务对象格式，通过可用组件上的大规模并行性获得其价值。为满足您的交互式导航需求，您可以继续挑选来源，清理该数据，以及将它保留在仓库中。但是可通过分析更多数据(可能甚至在最初似乎毫不相关的数据)来获取更多价值，以绘制所遇问题更可靠的情况。的确，数据可能在Hadoop中存在了很长时间，当您发现它的价值时，以及当它的价值得到证明并可持续时，就可以将其迁移到仓库中。

2. 为什么大数据至关重要

- 大数据解决方案是分析来自各种不同来源的原始结构化数据、半结构化和非结构化数据的理想选择。
- 需要分析所有或大部分数据而不只是一个数据抽样时，或者对一个数据抽样执行分析没有对更大的数据集进行分析更有效时，大数据解决方案是理想的选择。
- 大数据解决方案是在未预先确定数据的业务度量指标时，执行迭代式和探索式分析的理想选择。

谈到使用大数据技术解决信息管理挑战，我们建议您考虑以下问题：

- 传统分析模式的反向模式是否适合您遇到的业务任务？换句话说，您能否找到一个大数据平台可为您当前的分析工具提供补充并实现与现有解决方案的协调一致，以实现更好的业务成果？

例如，通常放在分析仓库中的数据必须经过清理、记录并且值得信赖，才能规范地放在严格的仓库模式中(当然，如果无法用传统的行和列格式存储它，它在大部分情况下甚至无法放在仓库中)。相反，大数据解决方案不仅会利用通常不适合传统仓库环境且数量庞大的数据，而且它将放弃数据的一些形式和“严格性”。好处是您可保留数据的真实性并能够访问海量的信息，在对信息采取您熟悉的适当行动之前探索和发现业务洞察；该数据可包含在一个循环的系统中，以充实仓库中的模型。

- 对于不能使用传统关系型数据库方法处理手头问题的方式来解决的信息挑战，大数据非常适合。

3. 为什么选择IBM解决大数据问题

您可以想象，IBM在集成解决方案方面拥有丰富的资产和经验，确保它们是兼容的、高度可用的、安全的和可恢复的，并且提供了一个框架，供数据在其整个信息供应链中流动，该框架是可信任的(因为没有人会只因为自己喜欢运行软件而购买一个IT解决方案)。

想想一个艺术家画一幅画时：一块空白的画布(一个IT解决方案)就是一个机会，您要画的图案是最终目标——您需要合适的画笔和颜色(有时您会混合一些颜色来使它完美)来画您的IT图案。对于只销售与服务或新进入市场的文件系统绑定的开源Hadoop解决方案的企业，讨论开始后会结束于将图画挂到墙上所需的锤子和钉子。您最终不得不走出去采购绘画用品，并依靠自己的艺术技巧来画这幅图画。IBM大数据平台就像是一个“数字色彩”绘画工具包，其中包括您所需的一切，能帮助您快速地框架，绘画，并悬挂一套充满活力的、详细的图画，以及您认为合适的任何自定义内容。在该工具包中，IBM提供您所需的一切，包括为开发、自定义、管理和数据可视化所设计的工具集，针对统计数据和文本预构建的高级分析工具包，以及Hadoop运行时的企业硬化，这一切都包括在一个自动化安装包内。

IBM世界级的、屡获殊荣的研究机构，继续通过高度抽象的查询语言、优化、文本和机器学习分析等接受和扩展Hadoop领域。利用开源技术的其他公司，特别是规模较小的公司，可能有大量项目(IBM公司也是如此)，但它们所具备的知识深度通常不足以了解对企业至关重要的特性集。例如，开源有文本分析和机器学习组件，但这些工具都还不完善，也不易于使用，并且其扩展性不如BigInsights中的工具，这一点对于企业而言真的非常重要。毫无疑问，对于某些客户而言，开源社区就是它们所需要的，并且IBM绝对尊重这个事实(这就是您可以从IBM单独购买一个Hadoop支持合同的原因)。对于希望获得传统支持和交付模型，并使用在文本和机器学习分析以及其他企业特性中数十亿美元投资的其他客户，IBM提供其大数据平台。IBM也提供其他优势供您考虑：

24×7直接工程师支持、以您的母语提供国际化的代码和服务等。我们实际上已拥有数千名员工可以与您配合，帮助您绘出自己的图画。此外，还有来自IBM的解决方案，如在BigInsights上运行的Cognos Consumer Insights，它可以推动您的大数据项目。

考虑到IBM在Hadoop系统上所添加的所有好处时，就可以理解为什么我们将BigInsights称为一个平台。在本章中，我们介绍有关IBM为大数据解决方案所带来的价值的非技术性细节。

4. 关于Hadoop: 大数据术语

一般认为Hadoop有两个部分：一个文件系统(Hadoop Distributed File System)和编程模式(MapReduce)。Hadoop中一个关键组件是内置在环境中的冗余性。不仅是数据冗余地存储在整个集群内的多个地方，编程模型也是这样，通过在集群中的多个服务器上运行程序的多个部分，可预期失败并自动解决这种问题。由于这种冗余性，我们可以实现在一个非常大型的商品组件集群中分发数据及其相关的编程内容。众所周知，商品硬件组件将失败(尤其是当您有非常多的商品硬件组件时)，但这种冗余性提供了容错，以及让Hadoop集群自愈的能力。这使得Hadoop可以跨廉价机器的大型集群向外扩展工作负载，以处理大数据问题。

Hadoop项目包括三部分：Hadoop Distributed File System (HDFS)、Hadoop MapReduce模型和Hadoop Common。要理解Hadoop，您必须理解文件系统的底层基础架构以及MapReduce编程模型。

MapReduce是Hadoop的心脏。正是这种编程模式，实现了跨越一个Hadoop集群中数百或数千台服务器的大规模扩展性。MapReduce概念对于那些熟悉集群向外扩展的数据处理解决方案的人来说相当易于理解。对于刚接触这个主题的人来说，它可能有些难以掌握，因为它不是人们以前一般接触过的某些概念。如果您刚接触Hadoop的MapReduce作业，别担心：我们打算以一种让您快速了解它的方式来形容它。

术语MapReduce实际上指的是Hadoop程序所执行的两个独立的、不同的任务。第一个是map作业，它拿出一组数据，并将它转换成另一组数据，其中每个元素都被分解成多个元组(键/值对)。reduce作业将map的输出作为输入，并将那些数据元组组合成较小的元组集。正如MapReduce这个名字的顺序所示，reduce作业始终在map作业后执行。

5. InfoSphere BigInsights: 分析静止数据

Hadoop在协助企业驾驭迄今难以管理和分析的数据方面提供了巨大的潜力。具体来说:

Hadoop能够利用各种结构(或根本不使用结构)处理海量数据。尽管如此，Hadoop从各方面讲仍是一项相当年轻的技术。Apache Hadoop顶级项目自2006年开始启动，虽然采用率在不断上升，并且越来越多的人参与开放源码编写，但Hadoop仍然存在不少人所共知的缺点(平心而论，即便是版本1.0情况也差不多)。从企业的角度而言，这些缺点可能会妨碍各家公司在生产环境中使用Hadoop，甚至可能会使它们拒绝采用Hadoop，因为客户往往会预期在生产过程中实现某些运营指标，如性能、管理功能和稳健性。例如,Hadoop分布式文件系统(HDFS)具有一个集中元数据存储(以下简称NameNode)，它表示一个会导致失去可用性的单点故障(SPOF)(版本0.21中增加了冷备用)。当NameNode恢复之后，可能还需要花费很长时间恢复Hadoop集群的正常运行，因为其跟踪的元数据必须加载至NameNode的内存结构，而所有内存结构均须重新构建和填充。此外，Hadoop结构复杂，难以安装、配置和管理，并且目前掌握Hadoop技术的人员还不是很多。同样，掌握MapReduce技术的开发人员资源也相当有限。编写在Hadoop环境下运行的传统分析算法(如统计或文本分析)难度很大，要求分析师精通Java编程，同时还能够熟练运用MapReduce技术开展分析算法(Pig和Jaql等高级语言简化了MapReduce编程过程，但仍需要经过学习)。内容还有很多，但您只需记住一点: Hadoop不仅需要一些企业强化，还需要加强工具和功能，使其能够帮助实现Hadoop平台提供的各种发展潜力(例如可视化、文本分析及图形管理工具)。

6. IBM InfoSphere Streams:分析移动数据

采用BigInsights后，我们将通过提供信息海洋保障您的竞争优势，IBM InfoSphere Streams (Streams)可帮助您监控流过环境的海量数据流。您可以深入挖掘数据流，为您的企业获取对时间敏感的竞争优势，也可以像身处海量数据流中的绝大多数用户一样只是充满敬畏地任由它们浩浩荡荡地流过。这就是Streams的用武之地。其设计可让您充分利用大规模并行处理(MPP)技术来分析数据，同时由于它不断流动，因而您还能实时监控发生的问题并采取行动、更加有效地做出决策，进而提高收益。

Streams是一个强大的分析计算平台，实现了实时分析数据(仅存在微小的延迟)。Streams不再像BigInsights一样搜集大量数据、操控数据、将数据存储到磁盘上，然后进行分析(换句话说，是指静止数据分析)，Streams可让您对移动数据运用分析技术。在Streams中，数据将会流过有能力操控数据流(每秒钟可能包含数百万个事件)的运算符，然后对这些数据执行动态分析。这项分析可触发大量事件，使企业利用即时的智能实时采取行动，最终改善业务成果。当数据流过这些分析组件后，Streams将提供运算符将数据存储至各个位置(包括BigInsights或其他数据仓库)，或者如果经过动态分析某些数据被视为毫无价值，则会丢弃这些数据(要么由于数据无意义，要么由于数据虽然存在特定用途，但要求持久性不强)。

如果已经对复杂事件处理(CEP)系统非常熟悉，您可能会在Streams中发现一些相似之处。不过，Streams的设计可扩展性更高，并且支持的数据流量也比其他系统多得多。此外，您还将了解Streams为何具有更高的企业级特性，包括高可用性、丰富的应用程序开发工具包和高级调度。

您可以将数据流比作一系列连结运算符。初始运算符(或单一运算符)通常是指Source运算符。这些运算符可读取输入数据流，然后反过来发送下游数据。中间步骤由执行特定操作的各种运算符组成。最后，每条进入动态分析平台的通道都有多个出口，并且在Streams中，这些输出内容被称为Sink运算符(就像水一样从水龙头喷涌而出流入厨房水槽)。

我们将Streams作为一个平台，因为您几乎能够以任何方式构建或自定义Streams，从而提供应用程序来解决各种业务问题。当然，它还是一个企业支持平台，因为这些运算符中的每一个均可在集群中的独立服务器上运行，从而提高可用性、可扩展性和性能。例如，Streams提供了丰富的工具环境，可帮助您设计流应用程序。Streams的另一个好处在于，它与BigInsights共享同一Text Analytics Toolkit，因而能够在整个大数据平台上实现技能和代码段重用。当您的流应用程序部署准备妥当后，Streams将在运行时根据集群类负载平衡和可用性指标自主确定处理元素(PE)的运行位置，从而使其能够重新配置运算符在其他服务器上运行，确保一旦服务器或软件发生故障，数据仍能持续流动。同时，您还能够以编程的方式指定哪些服务器上运行哪些运算符，并可在特定的服务器上运行流逻辑。

这种可自定义的自主式流平台只需增加额外的服务器和分配这些服务器上运行的运算符，即可增加执行数据流分析的服务器数量。Streams基础架构则负责确保数据在运算符之间的成功流动，无论运算符在不同的服务器上运行还是在同一服务器上运行，这样即可提供从最初建立小平台到根据需要不断发展平台所需的高度敏捷性和灵活性。

Streams不仅极其适用于结构化数据，而且也适用于其他80%的数据(包括传感器数据、语音、文本、视频、财务以及许多其他来源生成的非传统半结构化数据或非结构化数据)，这一点与BigInsights极为类似。最后，由于Streams和BigInsights均隶属于IBM大数据平台，您将会发现针对移动和静止大数据构建的分析技术均具有大量相同的高效功能。例如，从Text Analytic Toolkit构建的提取程序也可以在Streams或BigInsights中进行部署。

——*文字删节于Paul C. Zikopoulos, Chris Eaton, Dirk deRoos, Thomas Deutsch, George Lapis所著《理解大数据》(Understanding Big Data)一书*

IBM Corporation
Software Group
Route 100
Somers, NY 10589 U.S.A.

在中国印刷
2012年9月