

从数据库到数据空间，从服务于企业到服务于大众

——From Database to DataSpace, From Enterprise to People

孟小峰

网络与移动数据管理实验室

中国人民大学信息学院

引言

当代数据的三个典型特点使得传统关系数据库捉襟见肘、疲于应付。第一是海量，全球的数据量在以指数的趋势迅猛增长，据保守估计，目前每年全球至少将产生 15 亿 TB 的新数据产生。第二是共享，互联网和通讯设备的普及使人们享受在他人的数据带来的好处，数据库之间因此也建立起越来越密切的联系。第三是多样化，现在数据已不再是关系模型下纯粹的结构化的文本数据，图片、音频、视频乃至非结构化的文档都大量的涌入到人们的应用中来。数据库的研究者和制造商们并非无视这些事实，他们在功能和性能上仍在不断地丰富完善，不断地修补着这架越来越难以驾驭的马车。毕竟目前数据库在数据管理中仍旧占有主导地位，但这却不能表明它处在越来越尴尬的境地和最终会被代替的命运。其实上面提到的三个特点只是数据发展中表面现象，我们是根据这些特点继续维护，延续着“头疼治头、脚疼医脚”的这种治标不治本的补救措施？还是另辟蹊径，寻求一种新的数据管理技术在根本上进行大胆地变革？答案是不言而喻的，因为我们面临则者前所未有的新的变革和新的需求。

首先未来数据管理的主体将从单纯企业需求转向更为丰富的个人数据管理需求。

纽约时报著名的专栏作家托马斯·弗里曼在他的畅销书《世界是平的》一书中有这么一段话：

“我想全球划分为三个主要纪元。全球化 1.0 自 1492 年，持续到大约 1800 年。全球化 2.0 大概从 1800 年持续至 2000 年，中间曾经被大萧条及两次大战打断。2000 年世界进入了一个新纪元：全球化 3.0。世界从小缩成微小，竞赛场也铲平了。”

对此他又进一步解释道：

“在“1.0”，推动全球化的力量来自国家，在“2.0”，推动力来自企业，在“3.0”，推动力则来自个人。个人的力量大增，不但能直接进行全球合作，也能参与全球竞逐，利器即是软件，是各式各样的电脑程序，加上全球光纤网络的问世，使天涯若比邻。……”

从他的观点里我们可以得出，进入 21 世纪以后，随着个人电脑和互联网的普及，个人的影响力的提升使得在过去以企业为主导的模式逐渐地向以个人为主导的模式演变。

在过去的三十多年里，数据库技术主要服务于企业计算，我们几乎为企业的数据库管理开发了近乎完美的 DBMS。数据库作为当前最成熟的系统软件之一，已经成为了现代计算机信息系统和计算机应用系统的基础和核心。数据库也从最初的层次、网状数据库演变到了今天的关系数据库，为大家熟悉的 Oracle、DB2 和 SQL Server 等商业关系数据库已经广泛应用于各行各业。在很多人眼里看来，似乎一切都是如此的完美，所有的数据管理问题都会在这里得到答案。然而事实并非如此。

进入二十一世纪，我们忽然发现管理者世界上最大、最丰富的数据集，而且主要为个人服务的 Google, MSN, Yahoo 均不使用传统 DBMS，而是另辟蹊径去寻找能更好满足个人数据管理需要的方法。

不可否认数据库技术在过去三十年里为推动企业数据管理的发展所做出的无法替代的贡献，并将继续发挥其应有的作用。但在世界进入全球化 3.0 后，推动力正在由企业转变到个人，因此可以断定新的数据管理技术将由服务于企业的管理而过渡到个人的管理需求上面，那么数据管理技术将在服务于人的管理中起到什么样的核心作用？

其次新的计算机科学问题将是解决如何计算性能和计算成本不断改善，但人可用的时间和精力却恒定不变这一矛盾现象。

大家都知道计算机领域中的摩尔定律，它的一个广义的解读是这样的：计算机的性能随着时间呈指数级的增长；同时，计算的成本则会随时间呈指数级的下降。这一定律随着计算机领域的飞速发展得到了越来越有确凿的证明：CPU 的速度、内外存的容量在迅速增长，相对的是它们的价格却一路下跌。

摩尔定律及例外 Moore's Law and Exception

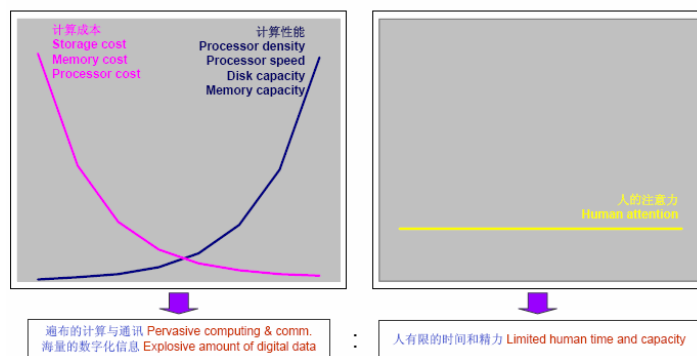


图 1.1 摩尔定律及例外

这样一个发展趋势的必然结果，就是计算和通讯会越来越普及，以至于数据量以难以想象的速度急剧膨胀，有人把这种现象称作是全球性的数据爆炸。也就是说目前计算技术的发展是使人更加应接不暇加速到来的信息，而没有丝毫减轻人们的负担。

据保守估计，目前每年全球至少将产生15亿TB的新数据。另一方面，在数据管理中却有一样东西是基本维持不变的——那就是人的注意力和人能够用在计算上的时间：每个人的总的寿命以及每一天用在工作中的时间在近千年中几乎没有太大的改变。于是作为数据管理技术的研究者和数据管理系统的使用者，我们发现正处于一对看起来很难调解的矛盾之中：一方面是遍布和汹涌而来的海量数据，另一方面则是人有限的时间和精力。这使得数据管理技术特别是传统的(关系)数据库管理技术面临越来越多的挑战，我们自己造的这块“巨石”已经压得我们喘不过气来。什么样的数据管理技术可以化解这对矛盾，如何使我们在场人与数据的大战中占得先机？

第三，数据是未来计算的核心。

我们计算机领域的人一直把速度作为计算的核心，所以孜孜不倦地追求提高计算机的速度和效率。正如 30 年前我们一直在抱怨天气预报、机器翻译的质量不好是因为计算机的性能不够好。然而事实是，到现在计算机的速度已经提升了上百万倍甚至上亿倍，但机器翻译并没有像预料的那样取得具体的突破性进展，天气预报一样该不准确还是不准确。这里有一个有趣的例子。在 2005 年的一次 NIST(美国国家标准与技术局)举办的机器自动翻译大赛中，最终结果让人大跌眼镜，冠军竟然是仅开始研究三年而且是首次参加的这次比赛的 Google。这个结果让该领域的专家们“伤心不已”。事实上，令 Google 获胜的“统计式”翻译算法，其基础是统计与分析某一单词在这一语言环境中被运用的概率与位置，来寻找词汇的排列规则；而另一种“很有前途”的热门算法，“类比式”算法，则是分析数以亿计的现成的翻译作品，

当需要翻译新的语句时，在现有的数据中搜索与之最相似的语句，来进行翻译——搜索和海量的数据分析，无论是哪一种，都是 Google 的专长。一句话，Google 制胜的法宝是其所多年积累的海量数据。这说明了计算的核心已不再是速度，而是数据，未来的世界承载在数据之上！但如果我们没有合适的的数据管理技术来使用这些海量而嘈杂的数据，对我们来说反而会是一场灾难。那究竟什么样的数据管理技术可以帮助我们驾驭这些数据呢？

归纳上述的三点我们不难看出，未来先进计算的核心是数据，而数据管理的主体不再是企业计算，而是围绕人的计算。着力解决人的时间和精力的问题将是我们面临的新的科学问题。

要解决这个问题，首先让我们看一下对应于个人管理现实中的数据，计算机中的数据管理主要面临的挑战：

- 数据的纷繁复杂。不管是结构化的还是非结构化，不管是文本，声音，图片，视频等等一切数据都是要管理的对象。
- 数据之间的逻辑关系。现实中之所以井然有序，在于人们对任何对象之间建立了逻辑关系，那么计算机中的数据如何进行关联是需要关心的重要问题。
- 数据的演化。现实中的对象会自我变化，会和其他的对象联合产生变化，那么对应计算机中的数据就是它们的自我演化问题。
- 数据的场合感应。要减轻人的时间和精力，就要能够主动地依其所处的场合（context）给出恰当的信息感应（awareness），减轻其处理数据的负担。
- 数据的可信性。解决人的管理问题，必须保证数据的真实，可信，和隐私保护等问题。

传统而严格的 DBMS 在这些现实面前将无能为力，这也促使我们去寻求一种新的数据管理技术，甚至可以进一步看作一种新的数据管理理念，这就是——数据空间（DataSpace）。基于以上分析，我们这里大胆地预测：未来数据管理技术将由数据库管理到数据空间管理，由服务于企业计算到服务于大众的计算。过去三四十年我们为企业打造了一个成功的软件，未来十年我们将为大众的需要创造造一个全新的软件。

数据空间——数据管理新概念

近年来，随着因特网的迅猛发展，web信息量急剧膨胀，日益成为一个巨大的数据库，对于这个实实在在的信息库，人们不知道其信息量的多少，不知道信息的存放位置，不知道信息的格式。这些海量的信息分布在世界各地的无数台计算机设备上，格式多样、内容丰富，有的与个人有关，有的与企业有关。这种数据信息存在方式的新特性，使人们对于数据资源的存储、访问等出现了新的特点：

- web日益成为一个信息闭包。web信息库中信息量的急剧增长，几乎包括了人们所需要任何信息，这一发展趋势对于传统的数据管理方法和技术形成了冲击，甚至改变了人们解决问题的思维方式，人们对于数据的价值和利用这种价值的方法有了新的认识。
- 传统的数据库技术不能满足新的数据管理的需要。Google、百度等网站对数据处理方式的改变也说明了这一特点，于是人们提出了一个新的概念：数据空间（DataSpace）。数据空间是存储个人、群组和企业信息的理想方式。

到底什么是数据空间，我们如何能够像传统的数据库一样，清晰的勾画出数据空间的内涵？

传统数据库的各种数据存储方式，关系也好，XML也好，无不强调一个格式，总是先有一个格式，然后使数据服从于这个格式，如此才能存储数据，进而提供查询等服务。但是任何形式的数据，其核心都是数据本身，形式只是一种载体，如果将数据限制于某种形式之中，多少显得有些被动，所以就是一种“被动”的方式，也就是说如果你有一份不同格式的数据要想存储于数据库中，必须将其转化为数据库中数据的存储格式。因此，对于这种格式性很强的存储，可以称之为“先有格式，后有数据”。

数据空间不同，从它的名字可以看出，它与数据库不同并且强调的是是一个 Space，Space 是什么？是空间，广阔的宇宙是一个 Space，是个 ObjectSpace，不管这些 Object 在其中如何排列，如何组织，只要是属于这个 Space 的就是符合要求的。同样，数据空间是一个满是数据的空间，数据在其中如何组织都可以，表也罢，XML 也罢，文本也罢，只要你是数据的一种载体，你就可以存在于这个 Space 中，对数据的组织排放不做任何要求，正如[9]中所说：“一个数据空间应该包含与某个组织或个体相关的一切信息，无论这些信息是以何种形式存储、存放于何处”。这样一来，无论你有一份怎样格式的数据，XML 文档也好，文本文档也好，都可以存储于数据空间中，并且通过数据空间来对其进行掌控，这可以称之为“淡化形式，凸现数据”。

这里简单概括一下数据空间的特性：

数据空间与实体相对应：数据空间是有所属的，与实体一一对应，一个人可以有一个数据空间，一个组可以有一个数据空间，一个企业可以有它的数据空间。数据项是数据空间的基本元素。数据空间是数据项的集合。数据项是与数据空间所对应实体相关的信息单元。一个数据项可以是一封电子邮件，保存下来的一个网页、一个文件等，也可以是一个传统数据库表。数据空间中的数据项一定是对于实体有意义的。有用性也是定义数据空间的边界，这种有用性可以是现实的，也可以是潜在的。

数据空间具有空间和时间特性：从空间上，数据空间的数据分布存放在许多位置；从时间上，数据空间中的数据也随着实体的发展而不断变化，一些新的数据项会加入进来，同时一些不再具有应用价值的项会消失。数据空间的大小是动态变化的，随着实体的进化，数据空间会不断进化，通过数据挖掘、自适应等技术，数据质量会不断提高，包含的信息量会不断增强。

实体数据空间交叉重叠：由于数据空间是与实体是对应的，不同实体所对应的数据空间是有重叠的，一个数据项可能即属于实体一的数据空间，又属于实体二的数据空间。

个人数据空间将是数据空间的主要存在和应用形式：在过去的30年中，数据库应用的主要对象是企业，可以预见，在未来的数据管理领域，数据管理将会转移到为人服务，为提高人的生活质量和效率服务，个人数据管理将是未来数据管理技术研究和应用的主要对象。相应地，个人数据空间也将是数据空间研究的主要对象。

综上，我们可以观察到数据空间本质上区别于传统的数据库（见表1）。

表1：数据空间与数据库的区别

数据空间	传统数据库
淡化形式，凸现数据	先有格式，后有数据
开放的，支持多种不同的数据源	支持有限的格式，封闭的
强调数据的可关联性和可演化性	关注数据的稳定性
具有 Pay-As-You-Go 特性	需要的时候集中建成
面向实体需要	面向应用主题

因此，数据空间将是一个开放的系统，其中包含与实体有关的各种数据，对其进行管理的目的是提高实体的运行效率。数据空间涉及很多技术，如数据的获取、组织、存储、索引；任务的管理；场合感应；隐私保护等等。

新的机遇与挑战

数据信息的新特点，使人们开始重新审视和定位数据空间技术。目前的情况与30年前的情况类似，那时企业的集中数据管理需求催生了数据库技术，现在对于基于web的个体数据管理需求期待着数据空间技术研究的重大突破。数据空间技术为研究者提供了重大机遇。

数据空间研究日趋活跃：目前，人们对数据空间相关技术的研究日趋活跃。Jens-Peter Dittrich Marcos, Antonio Vaz Salles等人将数据空间理论应用于个人信息管理，作了大量工作，2005年，在VLDB2005 发表了一篇Demo Paper: iMemex—将个人从信息枷锁中解放出来，实现了个人信息管理原型系统。其后进行了更深入的研究，分别在2006年VLDB和PIM Workshop发表了论文，系统阐述了个人数据空间管理的概念，提出了一个新的数据模型：iMEMEX，基于数据源视图的概念，建立了个人数据空间框架模型，并基于此实现了个人数据空间原型系统。

数据空间面临众多研究课题：尽管对于数据空间技术进行了一些研究，但是还不深入，在数据空间，还有很多挑战性的研究课题。

- 数据空间的理论基础

个人数据空间的研究涉及很多基础理论问题，如对数据空间中不同数据模型、数据关系和查询结果的理解，在传统数据库中，人们关注的是查询语言的表达能力，在具有上下文相关性的数据空间中，针对的是内容来自不同参与者的查询，这就面临许多问题，例如，我们如何检测具有不同语法结构但语义相同的查询。

- 数据空间中的数据关系和数据模型

传统的数据库管理系统大都是基于关系模型的，有完备的关系代数、关系演算理论。在数据空间中，传统的关系模型是否适用，网状模型和层次模型是否是更好的选择，相应于新的数据关系和数据模型特点，原来的理论是否适用；数据空间中也面临数据的一致性、安全性、并行性等问题，原来的并发处理策略、事务处理等理论是否适用；访问效率评价问题，和传统数据库不同，在数据空间中，影响访问效率的关键因素也不再是磁盘的I/O，那么这种情况下如何衡量访问效率，如何进行查询优化。

- 数据的存储和索引

面对如此巨大的个人数据量，在一台机器上进行数据的存储是不可能的，同时从安全性、可访问性等方面进行考虑，也不是很好的。而且Web数据是变动的，有的数据会随时消失，这样就要求我们对数据的安全性策略进行研究。

众多的研究课题为研究者提供了机遇。但是，与之结伴而来的是巨大的挑战。数据空间的挑战性来源于自身的特点。

- 由于数据空间的分布性，使得数据空间的数据组织、存储、索引、访问等都有新的特点。
- 数据空间中主体因素的作用。由于数据空间面向的实体千差万别，从而，依赖于实体的数据空间也有很多个性，因此对数据空间技术的研究，既要包括对通用数据空间模型和应用技术的研究，也要研究主体因素的作用。
- 数据的多样性给数据空间的数据输入带来挑战。数据类型愈来愈丰富，包括各种格式的电子文档、音频、视频、图像、移动数据等各种信息，数据空间必须建立一个

开放的能够适应各种数据格式的数据接口。

- 数据空间中的数据关系和任务管理。数据空间管理的目的是提高实体的效率，因此其不仅能够被动的显示主体所需要的信息，而且能够通过对数据关系的分析，自动的提醒主体应该做的的事情，这也是场合感应的基础，也是很有挑战性的研究课题。

此外，数据空间的实现模式是怎样的？如何评估数据空间管理系统的性能？等等，都是挑战性的问题。也正因为这些挑战的存在，愈发显现出它的魅力。相信随着研究的深入，数据空间及其管理技术，将成为新一代数据管理技术的核心，成为人们享受信息技术发展带来的巨大效益的基础平台。

综上所述，DataSpace 是一个“泛化”了的数据库，所‘泛’之处就在于数据形式的泛。这种泛化的 DB 从根本上来说还应该是 DB，只是相当于在传统 DB 之上又架设了一层，使得使用更加方便，但是同时也带来了挑战，正所谓便利多多，挑战多多。但是为了它所能带来的诱人便利，这些挑战还是值得一试的。