大数据技术发展的十个前沿方向（上）

Ten Fronties for Big Data Technologies (Part A)

吴甘沙，男，现任英特尔中国研究院院长。2000年加入英特尔，先后在编程系统实验室与嵌入式软件实验室承担了技术与管理职位，期间参与或主持的研究项目有受控运行时、XScale微架构、众核架构、数据并行编程及高生产率嵌入设备驱动程序开发工具等。2011年晋升为首席工程师，共同领导了公司的大数据中长期技术规划，主持大数据方面的研究，工作重点为大数据内存分析与数据货币化。在英特尔工作期间，发表了10余篇学术论文，有23项美国专利（10余项成为国际专利），14项专利进入审核期。

1 引言

“大数据”的发展与IT产业其他领域的发展相辅相成，近年来互联网、移动互联网、物联网、云计算和高性能计算等方面的高速发展从内涵上不断推动大数据的技术演进，从外延上不断延展大数据的应用范围。

多年来，笔者有幸接触国内外学术界和工业界的大数据研究，2014年底受清华数据科学研究院之邀，把所见、所得、所思总结为《大数据的十个技术前沿》的演讲。这次得到《大数据》杂志邀请，将其改为综述文章，并分为3期刊出，分别为：膨胀宇宙、巴别之难、数据有价；软硬兼施、多快好省、天下三分、分久必合；精益求精、人机消长、智能之争。笔者尝试从广度视角介绍大数据发展前沿的一些技术趋势和实践。限于篇幅，论述可能不够严密，介绍可能不够深入，唯愿抛砖引玉，激发同仁的思考和讨论。

2 十大前沿方向综述

大数据的根本出发点是指数思维方式。美国未来研究院（Institute of the Future）的发起人Roy Amara提出的Amara法则[1]认为人们往往会高估技术的短期影响力，而低估技术的长期影响力。数据总量的积累正是如此，在经历很长时间的缓慢增长之后，增长斜率会突然在一个临界点后急剧增加，变为爆炸式增长。人们常说：“最近两年产生的数据量相当于人类历史上产生的数据量总和的90%”、“现在产生的数据总量每两年翻一番”。所有这些桥段都指向同一个现象——指数增长效应。

在过去50年里，指数效应的主要驱动力是摩尔定律。英特尔的联合创始人之一戈登·摩尔（Gordon Moore）预言：每过18个月，晶体管数量翻一番，相应地中央处理器（central processing unit，CPU）性能翻番，成本折半，功耗折半。这种指数增长以链式反应的方式波及各个方面，如磁介质机械硬盘的容量增长以及主干网带宽的增长，甚至是每美元能够买到的数码相机的像素数都呈现了指数级的增长效应。最后，带来了数据的摩尔定律。

大数据发展的拐点已经到来，目前正在逐渐成为经济活动的主要承载者。数据被称为资产、原油、原材料、货币，无论哪种形容的方法都不过分。据IDC预测：2020年，70亿人的数据化生存以及500亿个互联设备的感知、互联和智能，将产生35 ZB的数据。1 ZB相当于1 000 EB，目前谷歌公司的数据量级为数十EB，这就意味着，一年将产生相当于1 000个谷歌公司的数据量。

从数据中提取出价值，海量数据才有存在的意义。大数据的生命周期和价值链条通常可以分成4个阶段：数据生成、获取、存储和分析。目前主流的大数据技术基本上是为了解决这4个问题。本文提到的10个技术前沿，基本上都落到这4个需求里，但总体来看又可以分成三大类。

- 解决数据本身的问题。分别为膨胀宇宙、巴别之难、数据有价。
- 解决大量的数据前提下，如何能够实时计算的问题。这里涉及技术手段与范式变迁，分为软硬兼施、多快好省、天下三分、分久必合。
- 分析如何能够提取更好、更精确的价值问题。分别为精益求精、人机消长、智能之争。

本期主要介绍膨胀宇宙、巴别之难、数

据有价3个技术前沿。

3 前沿方向一：膨胀宇宙

面对数据量的爆炸，IDC创造了一个名词——数据宇宙（data universe）。现在单机硬盘的容量已在TB级别，而商业公司的数据存储量级从PB到EB再到ZB，甚至再到下一步YB（美国国家安全局已经在犹他规划YB级别的数据中心）。与之对应的是存储技术的突飞猛进：存储介质技术发展、单服务器设计突破、分布式文件系统创新以及形形色色的分布式数据库爆发。

3.1 不断涌现的新存储介质

近年来，新的存储介质不断涌现，在性能和成本上都取得了长足的进步，构成了大数据发展的基础。

首先，磁介质的机械硬盘技术快速发展，单碟容量在TB级别翻倍增长。

其次，固态硬盘（solid state drives，SSD）获得了广泛普及，对革新存储体系结构起到画龙点睛的作用，例如SAP HANA[2]架构。又如AWS的SSD存储I2，Databricks用它在2014年的Daytona Gray类Sort Benchmark夺魁（并列）。

第三，PCIe SSD和闪存存储（flash storage）更为激进。从特立独行的Fusion-io到众望所归的NVMe，以其轻量级栈、低CPU开销、直接闪存访问带来高吞吐量和高IOPS（input/output operations per second，每秒进行读写（I/O）操作的次数）。

第四，包含闪存和磁盘的混合存储或联合存储是对软硬件协同设计的创新。谷歌公司的Janus智能地把数据在闪存和磁盘之间进行分配和迁移，闪存只存放1%的数据，却能服务28%的读操作。

第五，下一代非易失性随机访问存储器（non-volatile random access memory，NVRAM）也将渐渐走上舞台中央，它的特点包括访问性能接近动态随机存取存储器（dynamic random access memory，DRAM）（最短时延为DRAM的2倍）、容量大、数据不易失、字节寻址（闪存只能块访问）等。这些特性将改写整个存储体系结构的版图，必将带来内存空间和文件系统的融合。

第六，磁带在超大规模数据备份和管理上仍有一席之地。谷歌公司作为世界上最大的磁带机买家，利用磁带对EB级别的数据进行备份和管理，并通过位置隔离、应用层问题隔离、存储问题隔离、存储介质问题隔离等多种混合手段保证数据的可用性。

3.2 不断突出的单服务器的存储极限

在新存储介质层出不穷的同时，单服务器的存储极限也在不断突破。从2008年到2014年，主流单服务器内存从8 GB发展到现在的96～192 GB。货架产品里，单服务器最高内存容量可达48 TB。在硬盘方面，从2008年到2014年，主流单服务器磁盘容量从1 TB发展到48 TB。

2014年9月，英特尔开发者峰会展示了2U服务器可以容纳1.5 TB内存和100 TB硬盘，使高密度部署更上台阶。微软公司在同年10月份宣布推出的Azure G系列虚拟机，能够提供单虚拟机448 GB内存。这不但推动了大数据“内存计算[2,3]”的普及，而且模糊了内存和磁盘的边界，越来越多内存被用于缓存，甚至当成RAM Disk使用[4]。

3.3 创新的分布式文件系统

大数据技术的发展起始于分布式文

件系统(distributed file system, DFS)。当前,分布式文件系统以Apache HDFS为主,但用户需求在持续变化。一方面,数据中心的资源开始统一管理调度,分离的小集群被转换成统一的大集群,对存储系统的容量上限、存储的空间效率、访问控制和数据安全有了更高的要求。另一方面,存储系统的使用模式由周期性的批处理应用变成了交互式的查询和实时流式应用。

下面简单描述分布式文件系统的几个最新发展。

首先,HDFS(Hodoop Distributed File System, Hadoop分布式文件系统)新实现的HDFS缓存功能允许用户把某些常用数据块保留在堆外内存中,一方面可以增加数据带宽,减少时延;另一方面,可以用于不同应用之间的高速数据共享。

第二,支持分层的存储设备。数据中心一般都有内存、SSD和硬盘等存储设备,新型非易失存储器(nonvolatile memory, NVM)也呼之欲出,还有各类传统存储系统,如SAN(存储区域网络)、NAS(网络附属存储)和NETFS(网络文件系统)。因此,HDFS推出新功能heterogeneous storages(HDFS-2832)以支持异构的存储设备,适用不同应用的存储需求。

第三,加密文件系统。现在的典型部署是一个大集群容纳所有用户,由此带来的问题就是数据安全。HDFS的新功能——加密式文件系统(HADOOP-10150),使用AES-CTR加密算法,能够透明地对HDFS上的文件块加密、解密,并且只有很小的性能损失。

第四,内存文件系统,如RAMCloud[6]。它是由成千上万台普通服务器的主存组成的大规模存储系统,所有信息都存储在这些快速的DRAM中,内存取代了传统系统中的硬盘,而硬盘只作为备份使用。其目标是同时实现大规模(100~1 000 TB)和低时延(5~10 ms),比目前系统快100~1 000倍。在Spark[4]软件栈中也加入了内存文件系统Tachyon,特别适合迭代式的计算需求以及多应用共享数据。

最后值得一提的是纠删码(erasure coding),它最早应用于通信领域,通过编码机制实现传输过程中容错甚至纠错,如今它也被用到了大数据方向。英特尔公司和Cloudera公司一起推出了一种新的纠删码实现。

3.4 蓬勃发展的NoSQL数据库

同时,基于DFS技术和MapReduce技术的演进,发展出品类丰富的NoSQL数据库技术[3,6~12]。NoSQL数据库摒弃了关系模型的约束,弱化了一致性的要求,从而获得水平扩展能力,支持更大规模的数据。其模式自由(schema free),不再坚持SQL查询语言,因此催生了多种多样的数据库类型,目前被广为接受的如下。

(1)类表结构数据库

类表结构数据库是最早出现且在模式上也是最接近于传统数据库的NoSQL数据库,但多采用列存储。其源头是谷歌公司的BigTable[7],并且在此之上发展出HBase、Hypertable、Cassandra和着重安全的Accumulo(美国国家安全局使用)。

(2)文档数据库

数据保存载体是XML或JSON文件,从而能够支持灵活丰富的数据模型。一般文档数据库可以通过键值或内容进行查询。MongoDB是典型的文档数据库,也是DB Engines数据库排行榜中排名最前的NoSQL数据库(前10名当中只有两个NoSQL数据库,另一个是Cassandra)。

(3)键—值存储

因其易用性和普适性形成了NoSQL家族中最大的一支。键—值是最简单的一种

数据模型，在此之上可以实现更丰富的数据模型。目前，基于不同一致性和存储介质（内存、SSD或硬盘）形成了很多选择。比如，亚马逊Dynamo[9]以最终一致性为主，而Berkeley DB[10]则保证串行一致性；Memcached[11]和Redis是基于主内存的，而BigTable一族则是基于磁盘的。

除了上面3种数据库类型外，值得一提的是图数据库，将数据存储在高效的图结构中，典型代表是Neo4j。另一个案例，由谷歌公司工程师开发的开源图数据库Cayley针对Linked Data和图数据（如语义网络和社交网络）。

在NoSQL的蓬勃发展中，其重要理论支持“CAP（consistency，availability，partition tolerance）理论”也在演进。传统上CAP必须保证P（partition tolerance，分区容错性），而在C（consistency，一致性）、A（availability，可用性）中取舍。Eric Brewer在名为《CAP理论十二周年回顾：“规则”变了》[12]一文中指出：CAP理论的3选2这一结论太过简单化，实际情况要更复杂。首先，在同一数据中心，分区的情况很少出现，意味着在系统不存在分区的情况下未必要牺牲C或A；其次，C和A之间的取舍可以在同一系统内以非常细小的颗粒度反复发生，其取决于特定的操作、数据或用户；再者，这3种性质都不是非黑即白的，每个属性都有多种度量。在这个前提下，CAP理论的应用会更加复杂。Eric提出：CAP要在大部分时候允许完美的C和A；当分区存在或者可以感知时，需要定义一种策略来探知其存在，并根据CAP理论的指导对其进行处理。换句话说，创建一个CAP全都有的系统是可能的。

NoSQL一般损失强一致性以换取性能，而抽样方法允许用户牺牲精度，以加快大规模数据集上查询的响应速度。其代表为BlinkDB，主要思想包括两个方面：一个是自适应优化框架，从原始数据中建立和维护一个多维度的采样集合；另一个是动态采样策略，根据查询的精度和响应时间要求，决定采样数据的规模。在VLDB 2012的展示上，BlinkDB使用100个Amazon EC2节点组成的机群处理17 TB的数据，能够在2 s之内响应一系列的查询，速度是Hive的200倍，而错误率也被控制在2%~10%。

在NoSQL提出近4年后，来自The 451 Group的Matthew Aslett在2011年提出了NewSQL[13]数据库的概念。NewSQL既能提供近似NoSQL的性能和可扩展性，又能提供类似于传统关系数据库那样的关系模型、事务和SQL语言接口。从架构或者实现角度来看，NewSQL系统可以分成三大类。

（1）使用全新的架构

该类又可以分成两类：第一类系统一般使用shared-nothing（无共享）架构，所有的节点都具有处理事务的能力，系统具有近似线性的扩展能力，其可以是通用的数据库（如Google Spanner[3]）或者为某种特定场合设计的数据库（如VoltDB[14]）；第二类系统则使用主从架构，有专门的节点进行事务处理，这种设计使得系统的扩展能力会受到一定限制。

（2）各种MySQL存储引擎

MySQL是一个高度可扩展的架构，可以根据特定的应用场景为MySQL编写各种存储引擎，比较出名和成熟的有TokuDB、MemSQL、ScaleDB等。最新版本的MySQL 6.5既支持传统的关系数据模型，又支持键值对数据模型，此外还支持Memcached的访问协议。

（3）透明数据分区技术

与Cobar很相似，能够自动地对数据分区，并进行分布式事务管理，如dbShards、Scalearc和ScaleBase等。

作为NewSQL的一种主流，内存数

据库以其优越性能成为新宠，主要包括两类：一类是传统数据库加上内存选项，如Oracle 12c[15]（包括Exalytics和Exadata）、IBM DB2带BLU加速以及微软SQL Server 2014等；另一类是完全重起炉灶设计的新型数据库，包括Altibase、MemSQL、VoltDB、EXASOL、H20和SAP HANA等。不断增加的内存容量也为商业数据分析带来了新的可能：hybrid transaction/analytical processing（HTAP）在同一片内存中完成事务性的数据存取与分析过程，消除了数据ETL的代价。

Hadoop不支持ACID事务限制了其应用场景，如删除旧的记录、更新表格中任意一项等均无法在Hadoop生态圈的工具中完成。因此，Hadoop最新推出的特性也体现了NewSQL的影响。首先是Hive，从0.14版本开始能够在给定的限制下支持NewSQL操作；随后HBase也开始支持Transaction操作。

针对执行时间较长的操作，Hive推出了LLAP优化。其包括如下特性：有效降低启动开销；充分利用JIT优化引擎；对于向量算子采用多线程执行，并在这些线程之间共享元数据；异步I/O。这些优化与Tez等执行引擎相互独立，协同工作，以加快Hive的查询速度。被认为是Hadoop接班人的Spark也启动了称为Tungsten的项目，对Spark的核心引擎进行加速。Tungsten专注于改善Spark对内存和CPU的利用情况，主要包括以下3个改动：使用程序语义以改善JVM的对象模型和垃圾收集功能；设计cache-aware的算法和数据结构，以更好地利用层次存储体系（memory hierarchy）；利用代码生成（code generation），以更好地发挥现代编译器和CPU的能力。

谷歌公司仍然推动着超大规模广域数据库研究的前沿，连续推出Metastore、Spanner和F1。尤其值得一提的是Spanner，可扩展到几百万个机器节点，跨越成百上千个数据中心，具备几万亿个数据库行的规模。在最高抽象层面，Spanner就是一个数据库，把数据分片存储在许多Paxos状态机上，这些机器位于遍布全球的数据中心内，通过复制技术实现全球可用性和地理局部性，保证即使面对大范围的自然灾害时数据依然可用（它的开源克隆CockroachDB名字取自蟑螂，寓指其超强的生存能力）。与Spanner同时现身的是新一代的谷歌文件系统Colossus，它们将取代BigTable和上一代谷歌文件系统的核心地位。F1是建筑在Spanner之上的关系数据库。在上述的NoSQL/NewSQL数据库上衍生出很多针对特定用途的数据库。如OpenTSDB和KairosDB是基于HBase和Cassandra的时间序列数据库。

传统上，比较“小众”的科学计算数据库也开始向大数据融合，主要体现为并行数组数据库（array DBMS）。目前得到最多关注的是SciDB，其作为开源的科学领域数据库，设计初衷旨在提供多维数据管理，更好地支持具有科学计算特点的分析，比如它使用数组数据模型，允许行列交换，支持查询语言和数学计算，性能上比传统RDBMS快两个数量级。另一个相关工作是TileDB，作为一个针对数组数据做优化分块（tiling）策略的存储管理器，也将发展成为完整的分布式DBMS。它针对物理世界数据的高度skew和稀疏性，实现了非规则分块的策略，从而达到更高效的存储和负载均衡。

4 前沿方向二：巴别之难

圣经里有一个巴别寓言：在人类文明

初期，曾经是“天下人用同一种口音语言说话”，人类语言相同，因而能够高效地合作。于是他们聚在一起要造“一座城和一座塔，塔顶通天”。但是，神不容许人类破坏神所定的纲纪，所以一夜之间扰乱了人类的口音和语言，让人类沟通困难，最终放弃建造工程，从而分布到不同的地方去。那个城叫巴别城，塔叫巴别塔。自此以后，“大一统”成为人类的梦想，但是语言障碍是最大的阻碍。

数据世界也面临同样的问题。不同来源、不同地方的数据用不同语言（格式）表示，即使相同格式，其语意和度量衡也可能不同。这些因素极大地阻碍了数据共享，限制了数据使用的范围。另一方面，数据可能是不完备的，甚至是相互之间矛盾的，这样导致了一个问题，即没有办法利用更多的数据产生更好的价值。

为解决这些问题，Data Curation[1]应运而生，中文可译为“数据治理”。其原意是指在科学计算中的数据抽取、转换、保存和复用。后来逐渐扩展，数据治理包含在科学、人文、社会、教育所有领域，对数据进行发现、获取、质保、增值、重用的活动。在这里强调的是数据治理中与数据分享相关的技术——data munging / data wrangling（数据再加工）[2]。数据再加工是指把数据从原始格式中抽取出来，然后向其他格式转化的过程。以前这个过程以手工为主，现在将逐渐变为半自动和自动过程。这是一个很难的题目，参考NP困难的提法，将其称为DB困难。

数据再治理技术希望打破数据的语义隔阂。新科图灵奖得主Michael Stonebraker目前就在做data wrangling。他的goby.com项目（如图1所示），根据某些条件返回与suicide six相关的几个选项，如何甄别这几个选项是否代表着同一个东西。Stonebraker开发的Data Tamer系统能够模拟人的推理思路，从不同的选项里面发现不同的线索。首先比较这些选项的源网站，接着进入选项所指的网页，分析数据

1 https://www.ideals.illinois.edu/handle/2142/3493

2 http://www.quora.com/What-is-data-munging

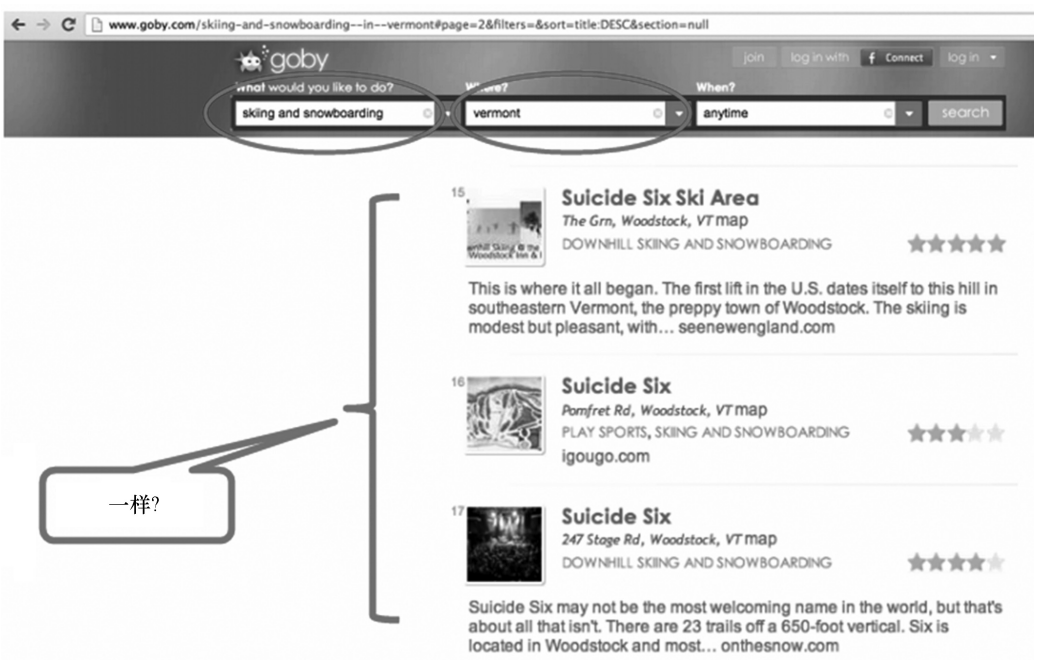

图1 数据发现示例

的异同。通过对数据进一步发掘，发现数据描述的主体有很多特征，以这些这些特征为基础，发现相似特征。通过证据的不断叠加，发现数据与数据之间的关联性。

Data Tamer技术的关键在于通过自动化的学习方式，发现数据中的规律和关联。首先是在文本这种典型的非结构化数据中发现结构；其次是发现重要的实体（entity）。而这一切都希望能够通过自动化学习来完成。同样在这个领域发力的还有Trifacta，该公司提出了“live in visualizations, not code”的口号，致力于让用户通过可视化完成data wrangling的工作。其基础是专门针对data wrangling任务设计的DSL，追求灵活和扩展的用户也可以在Trifacta提供的DSL上编写自己的脚本。

Data Wrangling下一步希望从半结构化或者多结构化的数据进一步扩展到完全非结构化的数据，如图片和语音。

数据治理完毕和数据质量提升以后，就是数据组织问题。

在今天的许多商业场景下，传统数据库和数据仓库在数据治理上暴露出难以操作和缺乏弹性的缺点。Schroeder认为Data Agility的重要性将不断上升，其关键在于组织数据。数据组织的复杂性使得数据很难被及时利用，更遑论进行实时更新，这极大地提高了数据使用成本[3]。

主流的大数据处理框架纷纷提升其数据描述和组织的灵活性。Spark在1.4版本中引入了称为DataFrame的新API。一个DataFrame就是许多列数据的集合，每一列都是被命名的。可以将其看作结构化数据中的表格或R/Python中的data frame，不同之处在于其支持许多优化算子。DataFrame可由多种来源构成，如结构化数据文件、Hive表格、外部数据库或者RDD结构。而GraphLab在图数据之外，也开始支持表结构SFrame。

另一个值得关注的数据组织工具是Apache的UIMA。IBM Watson在知识竞赛jeopardy中战胜了两个此项目的前世界冠军，其组织多种形态数据的基础就是UIMA，它的优点是组织数据以便于后期的分析。

5 前沿方向三：数据有价

数据作为未来经济的石油，自身必须有一个特性——价值。

数据的物理实质是记录在介质上的比特。比特是可以低成本无限复制的，这就和物品稀缺性矛盾了。物品失去了稀缺性后，其价值也就趋近于零。所以，数据有价首先要确保数据的权利。

为了确保数据的权利，先要保证数据的安全。大数据的安全本身又分为大数据系统的安全、数据本身的安全以及数据使用中的安全。

有了数据权利和保障数据权利的数据安全，数据才能进行定价。

5.1 数据权利

在互联网和物联网时代，数据的存在形式已经变得非常复杂。在整个价值链条中，有数据源头、数据收集者、数据存储者、数据使用者等。在多数商业场景下，他们都是不同客体。所以整个价值链中，权利的定义是一个重要的技术、商业和法律的课题。

如图2所示，笔者初步认为有如下5个基本权利。

- 拥有权。必须明确数据的拥有权，像其他的物理财产一样，拥有权可以出现变更和分割。
- 数据隐私权。即明确什么数据能够披露、什么数据不能披露、披露到什么样

3 http://www.cio.com/article/2862014/big-data/5-big-data-technology-predictions-for-2015.html

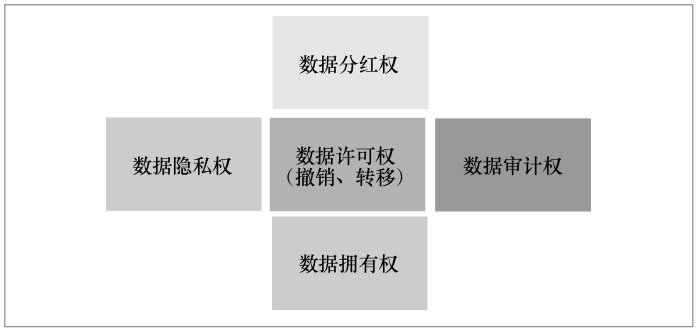

图2 数据的权利

的粒度。

- 数据许可权。哪些人在什么时间有权利看数据，是有约束的，比如今天允许给某个人看数据，明天就不允许。这个权利是可撤销的，也是可转移的。
- 数据审计权。监督用户按照某个规范许可使用数据。需要有一种审计机制，确保用户按照约定的许可规范使用数据。
- 数据分红权。基于数据外部性，获得数据使用许可的一方在反复使用数据中会产生新的价值，那么数据拥有者有没有可能得到分红？

5.2 数据安全

保障数据权利的核心是数据的安全问题。既有传统的信息系统安全问题，也有复杂的数据内容安全问题。

信息系统安全主要是大数据系统的安全控制，正在迅速地发展成熟。以Hadoop为例，加入了基于Kerberos的用户和服务鉴权、HDFS文件和数据块权限控制。比如Apache Accumulo是一个开源数据库，美国国家安全局几十个PB的数据存在这里，它采用了一种基于标签（label）的非常灵活的访问控制机制。在HBase里面也利用coprocessor的机制实现了类似的访问控制。

数据内容安全超越了访问控制和数据加密，更加复杂，可以称为“动态数据安全”。动态数据安全是大数据安全特有的新问题。

动态数据安全产生的原因是在监控和审计数据使用的过程中，不能简单地使用“允许/不允许”的静态策略来管理数据访问。数据一定要能被访问，否则数据就不能流动。关键是要在数据被访问和被加工的过程中动态地对数据流动方向、数据使用范围、数据使用粒度进行跟踪和监控。

数据监控主要分以下几个步骤完成。

（1）在数据产生的源头进行监控和规划

首先，个人对数据的控制。现在个人用户对自己的数据有了一定的控制能力，比如do not track功能可以防止互联网服务商根据cookie不断地跟踪用户行为，可以避免广告的retargeting，比如在京东商城上看中一双鞋，到了淘宝上它的广告还是跟着消费者这种情况。

另外，个人数据的删除。目前可以要求一些互联网的服务提供商把个人的数据删掉。值得一提的是MIT的创新项目OpenPDS（open personal data store），允许个人对自己的数据进行收集和控制，在保护隐私的前提下向第三方提供数据，并且获得价值。

（2）对数据分享的粒度进行控制

数据脱敏或匿名化是目前数据安全中最热的一个研究领域。如何保证开放数据里不泄露个人的隐私信息，是一个重大课题。在历史上很多的数据开放都导致了这样的问题。比如美国在线开放的匿名搜索数据，有人把这个跟美国选举公开信息进行了匹配，使得某些个人的隐私被暴露出来。

传统的脱敏方法是去标识符。比如一张表有姓名、年龄、性别、邮编和疾病几列，姓名是可以唯一标识个人的，叫做标识符。针对隐私的攻击方式还有很多。比如多数据源的相互匹配，Netflix尝试在去标识后开放了一些数据，但是有人把去标识后

的数据跟IMDB做了匹配，把一些有同性恋倾向的人找了出来，这就是多数据源的攻击。而研究表明，根据年龄、性别和邮编的信息，有90%以上的概率可以定位个人，这些属性叫准标识符，而这种攻击基于数据概率分布。

要防止这些隐私攻击，现在推出了很多技术，如*K*-anonymity[16]。*K*的意思是在所有准标识符都相同的组别里（比如，在上述的数据表例子中，年龄、性别和邮编都相同的所有数据记录）保证至少有*k*个相同的记录，从而提高单个记录被多数据源交叉定位的难度。后续发展出了*L*-diversity[17]和*T*-Closeness[18]，继续对跨组别敏感信息的统计分布提出更高的可区分度的要求。

2006年提出的差分隐私（differential privacy）[8]是近几年最热门的匿名化方法。这项技术提出，在数据中人为地插入噪声，同时通过精确模型设定保证噪音的程度不足以干扰各种数据分析算法（已经实际展示的有决策树、分类、聚类等），这样可以实现数据价值（信息粒度）和数据安全的平衡。

（3）建立数据使用的安全框架

未来，数据使用能够做到可用但不可见，相交但不相识。因为在几乎所有的大数据场景下，真正重要的数据分析结果，其实原始数据不是必须被公开或者传递的。为了实现这个目标，牵扯到以下几种技术。

- 同态加密。典型的是CryptDB/Monomi[19]，能够在加密的数据库上运行正常的SQL查询，而不用担心数据的明文被泄露，谷歌、SAP等公司都采用或借鉴了CryptDB的技术。

- 基于加密协议的多方安全计算。图灵奖得主姚期智先生1982年开始研究这个问题，叫做“百万富翁的窘境”：两个百万富翁要比谁更有钱，但是谁都不愿意说出自己的财富数值，这就是典型一种保护隐私下的多方安全计算场景。

- 基于可信计算环境的多方安全计算。前两种需要涉及晦涩难懂的加密算法，而基于可信计算环境的多方安全计算对数据计算的改变最小，也最有前途。当然，可信计算环境需要一些硬件支持。英特尔平台上开发了TXT、TPM、VT-d，目的都是保证应用计算环境是可信、可溯源的，计算中的数据被隔离保护。即将推出的下一个技术叫SGX[20]，它保证数据在磁盘和内存里面都是加密的，只有载入CPU里面进行计算的时候才是明文，更进一步隔离了磁盘和内存的物理攻击机会。

（4）区块链与零知识证明、多方安全计算等融合

在未来高度分布、去中心化场景下，可能会发展出各个数据实体之间不存在单个核心节点的安全控制机制。最典型的就是比特币所依赖的区块链（block chain）[4]技术被广泛看好，将承担全球规模的去中心化金融系统中事务记录、支付、数据资产管理和交易、智能合约等业务，以太坊（Ethereum）[5]是实现这些业务的开放应用开发环境。区块链技术也将被应用于个人数据控制（如上述OpenPDS的下一代Open Mustard Seed框架）和分布式数据存储（如MaidSafe）。区块链与零知识证明、多方安全计算等融合，将有可能成为下一代互联网基础设施平台。

5.3 数据审计监管的技术

系统安全、数据安全、使用安全都需要审计作为保证。所谓审计就是给出一个数据使用的条款，按照条款监控数据的使用。设计条款必须有形式化的描述，其目的在于让非IT的专业领域人员编写这些条款，如企业法务。如果一个企业的数据要

4 http://www.bitcoin.org/bitcoin.pdf, 2012

5 http://ethereum.org/ethereum.html, 2013

开放给另外一个企业，需要法律人士给出逻辑严格的使用条例，条例的内容本质上不是IT范畴。同时，因为条例规范是形式化的，IT技术方案也可以据此对数据的使用进行必要的审计监控。

5.4 数据定价的技术

数据定价是最具挑战性的研究方向，尚无成熟的研究成果。目前数据的定价有两个依据：一是根据效用，二是根据稀缺性。数据效用简单来说，就是数据使用的频率，也可以理解为从分析结果逆推数据的渊源（lineage），从而量化各方数据对结果的贡献度。稀缺性则是根据数据价值的密度以及历史价格的稀缺性进行定价。

5.5 数据咖啡馆

基于上述这些前沿技术，英特尔中国研究院开发了一个数据分享原型平台——数据咖啡馆。咖啡馆的寓意是让不同的人能够聚在一起进行思想的碰撞，产生新的价值。数据咖啡馆希望能够让不同方的数据碰在一起，产生新的价值。

许多独立垂直电商或者线下行业用户，仅靠其自身收集的消费者数据不足以对消费者建立精准的营销模型。因此，他们需要彼此间开放数据，甚至从通信、地图等专业数据源持续地购买数据服务。

另一个案例是癌症的研究和治疗。癌症是一个长尾病症，过去50年癌症的治愈率只提升了8%，在所有的疑难杂症中是提升最少的，很大的原因是不同研究机构癌症的基因组样本非常有限。但是，共享基因组样本受到严格的隐私法律的限制。英特尔中国研究院希望通过技术创新把这些数据汇聚到一起，加速癌症研究的技术突破。现在，英特尔中国研究院跟美国几家研究机构有一个愿景：在2020年前，一天之内一个癌症患者来到医院能够完成全基因组测序，同时分析出致癌的基因，并且给出个性化的治疗方案。

数据咖啡馆的目标就是帮助这样的场景能够持续、高效、低成本地运作。其创新点包括：集成了分布式云环境下的可信任大数据计算环境；形式化地描述数据使用规范；探索基于数据使用规范的程序检查器，包括对代码的静态检查以及对结果的动态检查。

未来数据咖啡馆的应用场景：企业的数据拥有方是一方，但是没有分析能力；具有分析能力的独立的数据使用者又是一方。数据拥有方的IT人员准备了数据存储和数据格式，商务和法务人员编写数据使用规范。将数据格式和数据使用规范提交到数据咖啡馆云。数据使用方的分析师们编写分析代码，并提交到云上。云首先对代码进行检查，把它拆成预处理和全局分析两部分，其中预处理部分在数据拥有方的防火墙内执行。发送前，在云内先运行一个静态的检查器，根据数据使用规范检查代码的合法性。只有通过合法性检查的代码才会被送到数据提供方进行计算。然后，把阶段性的处理结果送回云。在送回前，由动态检查器对结果进行审计检查。只有完全符合数据使用规范，全局分析部分才能收到预处理结果，并在云里面完成最后的计算。这个架构可以自然地衍生到多方的数据计算。

这个架构创新点在于：数据的提供方和数据使用方实现了可控的隔离。原始数据和核心分析算法作为参与各方的核心资产，在计算过程中得到保护，并且计算过程不受保护措施干扰。英特尔中国研究院愿意与各位同仁在这一领域共同开展前沿研究。

参考文献

[1] Amara R, Lipinski A J. Business Planning for AnUncertain Future: Scenarios & Strategies. New York: Pergamon Press, 1983

[2] Färber F, Cha S K, Primsch J, *et al*. SAP HANA database: data management for modern business applications. ACM Sigmod Record, 2012, 40(4): 45~51

[3] Corbett J C, Dean J, Epstein M, *et al*. Spanner: Google's globally distributed database. ACM Transactions on Computer Systems, 2013, 31(3)

[4] Zaharia M, Chowdhury M, Das T, *et al*. Resilient distributed datasets: a fault-tolerant abstraction for in-memory cluster computing. Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation, San Jose, CA, USA, 2012

[5] Li H, Ghodsi A, Zaharia M, *et al*. Tachyon: reliable, memory speed storage for cluster computing frameworks.Proceedings of the ACM Symposium on Cloud Computing, Seattle, Washington, USA, 2014: 1~15

[6] Ousterhout J, Agrawal P, Erickson D, *et al*. The case for RAMClouds: scalable high-performance storage entirely in DRAM. ACM SIGOPS Operating Systems Review, 2010, 43(4): 92~105

[7] Chang F, Dean J, Ghemawat S, *et al*. Bigtable: a distributed storage system for structured data. ACM Transactions on Computer Systems, 2008, 26(2)

[8] Dwork, Cynthia. Encyclopedia of Cryptography and Security. New York: Springer US, 2011

[9] DeCandia G, Hastorun D, Madan J, *et al*. Dynamo: amazon's highly available key-value store.ACM SIGOPS Operating Systems Review,2007, 41(6)

[10] OlsonM A, Keith B, Seltzer M I.Berkeley DB. Proceedings of USENIX Annual Technical Conference,Monterey, CA, USA, 1999

[11] Jose J, Subramoni H, Luo M, *et al*. Memcached design on high performance rdma capable interconnects. Proceeding of IEEE International Conference on Parallel Processing (ICPP), Taipei,China, 2011

[12] Brewer E. CAP twelve years later: how the "rules" have changed. Computer, 2012, 45(2): 23~29

[13] Moniruzzaman A B M. NewSQL: towards next-generation scalable RDBMS for online transaction processing (OLTP) for big data management. arXiv Preprint, 2014, arXiv:1411.7343

[14] Stonebraker M, Weisberg A. The VoltD Bmain memory DBMS. IEEE Data Engineering Bulletin, 2013, 36(2): 21~27

[15] Greenwald R, Stackowiak R, Stern J. Oracle Essentials: Oracle Database 12c. Sebastopol: O' Reilly Media Inc, 2013

[16] Sweeney L. K-anonymity: a model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557~570

[17] Machanavajjhala A, Kifer D, Gehrke J, *et al*. l-diversity: privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data, 2007, 1(1)

[18] Li N H, Li T C, Venkatasubramanian S. T-closeness: privacy beyond k-anonymity and L-diversity. Proceedings of the 23rd International Conference on Data Engineering, Istanbul, Turkey,2007

[19] Popa R A, Redfield C M S, Zeldovich N,*et al*. CryptDB: protecting confidentiality with encrypted query processing.Proceedings of the 23rd ACM Symposium on Operating Systems Principles,Cascais, Portugal, 2011

[20] McKeen F,Alexandrovich L, Berenzon A, *et al*. Innovative instructions and software model for isolated execution. Proceedings of the 2nd International Workshop on Hardware and Architectural Support for Security and Privacy, New York, NY, USA, 2013 □