IT@Intel 白皮书

英特尔 IT 部门
IT 最佳实践
云计算和信息安全
2012 年 6 月

利用数据匿名化（Data Anonymization）技术增强云的信息安全

尽管在数据匿名化能够实施在实际计算环境之前还需要多做研究工作，但此技术可以缓解一些信息安全的问题，简化隔离区和安全配置，并提高云计算的安全。

总体概述

英特尔 IT 部门正在探索数据匿名化技术 — 将已发布的数据弄得模糊来防止识别关键信息 — 以支持我们的混合云计算模式愿景和员工和客户隐私保护要求。我们坚信数据匿名化对于提高云计算的安全是一种切实可行的技术。

虽然我们知道无法实现 100% 安全的云基础设施，但是我们正在探究是否可以使用匿名化数据来提高云基础设施的安全。通过数据匿名化，数据对于其他人将变得毫无价值，但是英特尔 IT 部门仍然可以有效地处理这些数据。

我们执行了概念验证（proof of concept，简称 PoC），使用数据匿名化保护存储在公共云中的事件日志数据。PoC 成功地证明了数据匿名化非常奏效，而且被模糊化的数据仍然可以用于分析。我们可以基于匿名化的数据执行性能和安全分析。

- 我们从性能分析中发现了一些性能方面的问题，例如：某个网站因为在用户访问实际内容之前执行了两次重定向，从而增加了访问时间。
- 我们从安全分析中检测到一些真实的安全事件，例如网络服务器上出现探测器（probe）。

尽管在数据匿名化在实际计算环境使用之前还需要做更多的研究工作，但是此技术可以缓解一些信息安全的问题，简化隔离区和安全配置，并提高云计算的安全。

我们计划进一步探索数据匿名化，其中包括执行更加广泛的概念验证（PoC），开发更多的数据匿名化使用案例，让企业云用户了解数据匿名化的潜在优势和劣势，以及记录现有的开源数据匿名化应用。

Jeff Sedayao
英特尔 IT 部门企业架构师

目录

总体概述 1

背景 2

数据匿名化概念和技术 2

使用匿名化保护隐私 3

概念验证 6

PoC 实施 6

成效 7

主要经验 7

下一步工作 8

总结 8

缩写词 8

IT@INTEL

IT@Intel 计划将全球各地的 IT 专业人员及我们机构中的 IT 同仁紧密联系在一起，共同分享经验教训、方法和战略。我们的目标十分简单：分享英特尔 IT 部门最佳实践，创造业务价值并实现 IT 竞争优势。如欲了解更多信息，请访问 www.intel.com/cn/IT 或联系您当地的英特尔代表。

背景

云计算不但有助于降低成本、提高业务灵活，而且还能够使 IT 部门专注于高回报的投资项目。英特尔已从实施企业私有云中获益匪浅，并且其 IT 部门在混合云使用模式开发方面取得了显著进展。在追求混合云这一愿景之际，我们同时也意识到有必要根据英特尔信息安全策略保护数据。

对于英特尔 IT 部门乃至整个行业来说，安全和隐私问题都是阻止公共云广泛部署的一大障碍。IT 机构通常不愿意将价值高的机密数据存放在他们无法掌控的基础设施中。这一点在欧洲表现尤为显著，在欧洲对于个人身份识别信息的使用和保护有着严格的法律规定。

此外，随着统一计算设备的不断增长，越来越多的设备配备了全球定位系统，具备客户端感知功能、基于位置的社交媒体应用也变得日益普遍。在这些情形中，位置跟踪成了一个问题 — 位置信息对于提供定制和本地化服务非常有帮助，但是位置数据的存储和挖掘涉及到了隐私和监管问题。共享云基础设施的多租户可能会带来额外的问题，即故意或意外泄露数据。

通过积极应对潜在的安全漏洞，我们正在努力探索各种方法以提高云中数据的安全。其中一个方法是以一种可使云数据对于其主人以外的其他人来说毫无用处的方法存储云数据。

所有形式的数据保护都涉及到信息安全与简易实施之间的平衡问题，而要想获得更加安全的系统则需要我们付出最大的努力。

例如，同态加密（homomorphic encryption）— 一种强加密方法，支持无需输入解密的计算 — 就是一种可行的解决方案。同态加密尽管从管理论上是可行的办法，但是因为计算量太大而不适合实际使用。

一般来说，要想在隐私问题保护方面取得进步则需要在处理已发布数据方面做更多工作。我们需要找到一款切实可行的云数据存储方案，即能够保护隐私，又易于实施。

数据匿名化概念和技术

匿名化是一种技术，企业可以用它来提高数据在公共云中的信息安全，同时还支持分析和使用数据。数据匿名化把将要使用的数据以某种方式更改或发布，防止识别关键信息。

通过数据匿名化，机密数据的关键部分将被模糊化，从而保护了数据隐私。但是该数据仍然可以被处理以获得一些有用信息。匿名化的数据可在云中存储并处理，而无需担心其他人捕获该数据。稍后，结果将收集并映射到安全区域的原始数据中。

使用匿名化保护隐私

图 1 展示了使用数据匿名化保护机密数据的一个简单示例（从理论上来讲）。在该示例中，我们的目标是在不披露几家公司名称的情况下计算它们的总收入。要实现该目标，需要更改公司名称。例如，在原始数据中公司 A 的名称在基于云的数据中将被更改为“Bob”。而且，虚构的记录也将被添加到基于云的数据中，以便进一步将数据匿名化。转换表将存储在一个安全区域（通常是企业网络中的一个安全区域），可用于映射公司名称解析和识别虚构数据。

使用匿名化数据，我们可以在云中计算总收入，并且不会暴露机密信息。使用转换表，我们可减去虚构的公司数量，内部获得正确的结果。

此类匿名化可以阻止某些类型的数据挖掘攻击，因为在添加虚构数据后，将无法确定公司的数量以及收入最高或最低的公司。

但是如果操作不当，匿名化将会造成严重后果。例如，一家流行在线电影供应商公布了一个用户和电影评选结果数据库。为了保护客户身份，供应商会用随机编号替换客户名称，并删除个人详细资料。安全研究人员发现，通过综合分析数据与互联网电影数据库中公开的信息，可以揭露许多个人用户的身份。[1] 这项研究使我们充分认识到数据匿名化不只是将数据库中的字段简单地删除。

几种形式化安全模式可以帮助改进数据匿名化，其中包括 k-匿名化和 l-多元化。下面两部分将详细介绍这些安全模式，并举例说明。

k-匿名化

k-匿名化是由 L. Sweeney 创建的一种形式化隐私模型。[2] 其目标是让试图识别数据的人们难以区分每项记录与界定数量（k）的其它记录。

如果将一组数据 k-匿名化，并且每项数据记录中都有一组预先设定的属性，那么至少有 k-1 个其它记录与这些属性匹配。例如，假定一个数据集包含两个属性：性别和出生日期。如果该数据集执行了 k-匿名化，对于每项记录来说，都有 k-1 个其它记录与其拥有相同的性别和出生日期。一般来说，k 的值越大，得到保护的隐私将越多。

k-匿名化将这些属性分配到数据属性中，并要求以特定的方式处理它们，如表 1 所示。

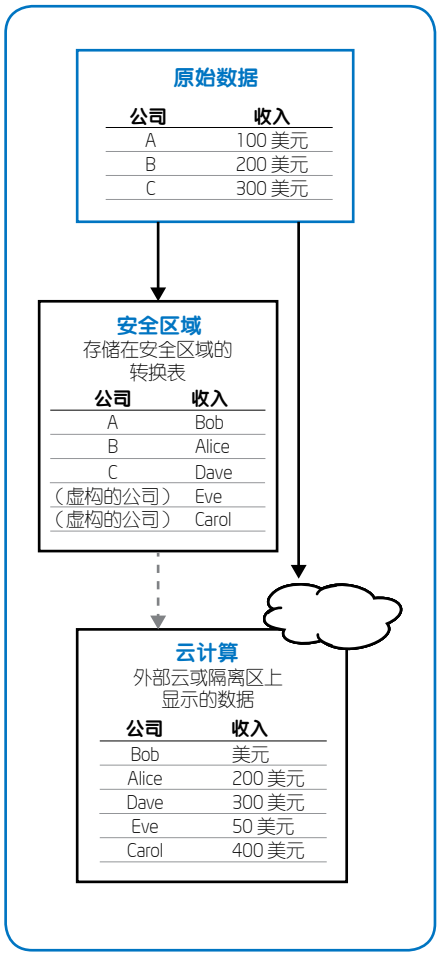

图 1. 数据匿名化有助于实现更加安全的云计算。

[1] Schneier，Bruce，《为什么“匿名”数据有时并没有匿名。》连线杂志（2007）。www.wired.com/politics/security/commentary/securitymatters/2007/12/securitymatters_1213

[2] Sweeney，L.，《k-匿名化：一种隐私保护模式。》*INT J UNCERTAIN FUZZ 期刊* 10（2002）：557-570. www.epic.org/privacy/reidentification/Sweeney_Article.pdf

表 1. k-匿名化属性

属性类型	属性	示例	所需操作
关键属性	可以直接识别个人身份	名称、社会保险号码	删除或模糊化
准标识符	结合外部信息可识别个人身份	邮政编码、出生日期和性别	隐藏或模糊处理
机密属性	个人感到机密不愿揭露的数据	收入、疾病类型	需要从个人资料中删除

表 2. k-匿名化病人档案样本

邮编	年龄	疾病
130▪	2▪	心脏病
130▪	2▪	心脏病
130▪	2▪	心脏病
130▪	2▪	病毒性感染
130▪	3▪	癌症
130▪	3▪	癌症

▪ 表示隐藏值。

表 3. 使用 l-多元化方法创建的病人档案，其中 l=2

邮编	年龄	疾病
130▪	2▪	心脏病
130▪	2▪	心脏病
130▪	2▪	心脏病
130▪	2▪	癌症
130▪	3▪	癌症
130▪	2▪	病毒性感染
130▪	2▪	病毒性感染
130▪	3▪	病毒性感染
130▪	3▪	病毒性感染
130▪	3▪	癌症
130▪	3▪	癌症

▪ 表示隐藏值。

研究表明，结合邮政编码、出生日期和性别，可在美国识别 87% 的人的身份。

一些机密属性，例如疾病类型等，应从个人资料中删除。为了保证 k-匿名化，数据中应该包含 k 个相同的准属性序列。

表 2 包含了一家假想医院中的病人档案样本：邮编和年龄是准标识符，而疾病是机密属性。邮政编码和病人年龄均被隐藏，表中所列的病人年龄仅表示其所处的年龄段。该表对于每个准标识符都提供了两个副本，从而实现了匿名化，其中 k=2。

k-匿名化可以保证人们无法从一组大小为 k 的数据集中给定数量的人群中识别个体。如果没有 k 个相同的准标识符序列，虚构记录可以被添加到数据中，在某些情况下必须消除这些虚构记录对数据处理产生的影响。

l-多元化

k-匿名化容易受到很多攻击。例如，使用表 2 中的数据，从理论上讲至少会遭遇两次攻击。

- **基于同质的攻击。**如果攻击者知道 Bob 在资料中有记录，并且年龄为 30 多岁，那么他将推断出 Bob 患有癌症。
- **背景知识攻击。**如果攻击者知道 21 岁的 Yuko 在资料中有记录，那么他将推断出 Yuko 患有病毒性感染疾病，因为一名年轻的妇女不太可能有心脏病。

使用另外一种隐私保护模式 l-多元化则可以应对上述两种攻击。l-多元化改善了匿名化技术，可提供 k-匿名化所不能提供的功效。两者的不同之处在于：k-匿名化要求每个准标识符组合具有 k 个条目，而 l-多元化则要每个准标识符组合中有 l 个不同的机密值。表 3 向表 2 中的数据添加了 l-多元化。该数据采用了 k-匿名化，其中 k=4，以及 l-多元化，其中 l=2。使用表 3 中的匿名化数据后，上文所述基于同质的攻击和背景知识攻击都不可能发生。

尽管 l-多元化较 k-匿名化增加了更多隐私，但是自然出现的机密属性可能因不能提供足够的种类而无法实现 l-多元化。将虚构的数据添加入资料中虽然会加大发生概率，但是却会为分析工作带来巨大麻烦。此外，概率推断仍然是可能的。例如，从表 3 中的数据可以确定 Yuko 患有癌症或者病毒性感染疾病，两者的概率是一半一半。

表 4. Anonymization Techniques to Obscure Data

技术	描述和应用	示例 IP 地址	名	姓	员工编号	工作年限	电话号码	月薪
		样本数据						
		143.183.23.3	Bob	Smith	325211	13	408-555-2935	5,000 美元
		143.183.23.10	Alice	Jones	452893	3	408-555-2931	4,000 美元
隐藏	▪ A 值被替换为一个常数值（通常为 0）；有时被称为“黑色标记”。 ▪ 有助于隐藏无需处理的机密属性。例如，无需公布的地址簿数据和薪水信息。	143.183.23.3	Bob	Smith	325211	13	408-555-2935	0
		143.183.23.10	Alice	Jones	452893	3	408-555-2931	0
		在薪水值中添加了 10,000 美元。						
散列	▪ 将每个值映射到一个新（不一定唯一）值中。 ▪ 有助于将一个较大的、可变的数据映射到具有一定长度的数据中。	143.183.23.3	683452343981113408			555	408-555-2935	5,000 美元
		143.183.23.10	23495103429323408			555	408-555-2931	4,000 美元
		名、姓和员工编号散列在一个固定长度的数字中。						
置换	▪ 将每个原始值映射到一个独特的新值中。 ▪ 如果安全区域中有转换表，则可以将新值转换回原始值。	143.183.23.3	Rob	Clemente	325211	13	408-555-2935	5,000 美元
		143.183.23.10	Eva	Gonzales	452893	3	408-555-2931	4,000 美元
		姓和名被映射到新值中。						
位移	▪ 在数值中添加一个固定偏移量。 ▪ 有助于隐藏数据，同时支持在云等环境中计算。	143.183.23.3	Bob	Smith	325211	13	408-555-2935	15,000 美元
		143.183.23.10	Alice	Jones	452893	3	408-555-2931	14,000 美元
		在薪水值中添加 10,000 美元。						
枚举	▪ 将每个原始值映射到一个新值中以保持排序。 ▪ 支持要求排列的数据分析，例如按薪资顺序。	143.183.23.3	Bob	Smith	325211	13	408-555-2935	25,000 美元
		143.183.23.10	Alice	Jones	452893	3	408-555-2931	2,0000 美元
		虽然更改了薪资数据，但是保留了相关的薪资排序。						
截断	▪ 缩短字段，删除末尾处的数据。 ▪ 有助于隐藏数据，同时仍可以保留数据信息。	143.183.23.3	Bob	Smith	325211	13	408	5,000 美元
		143.183.23.10	Alice	Jones	452893	3	408	4,000 美元
		区号后面的电话号码被截断。这虽然隐藏了电话号码，但是保留了员工的居住地信息。						
保留前缀	▪ 在 IP 地址中保留 n 位的前缀。	143.183.79.169	Bob	Smith	325211	13	408-555-2935	5,000 美元
		143.183.3.25	Alice	Jones	452893	3	408-555-2931	4,000 美元
		IP 地址虽然混乱，但是 16 位的前缀被保留下来。						

数据匿名化与加密的不同之处

尽管数据匿名化和加密是两个密切相关的话题，而且都能够帮助人们保护基于云的数据不受隐私和安全漏洞的攻击，但是它们并不是同一回事。

- 数据匿名化是一种数据转换技术，通过对数据进行有效的处理，从而使人们无法将数据与个人身份、目标或机构联系起来。
- 加密是指通过转换数据，使那些没有密钥解密的人无法读取这些数据。

加密是在匿名化实施中是一款非常有用的工具，尤其是当在一组数据中隐藏身份信息时。虽然加密非常有用，但是对于匿名化来说，既没必要也不充分。数据可以在不加密的情况下成功匿名化，但是加密的数据不一定需要匿名化。

示例

表 5 展示了一个示例文件，目标是根据位置和工作年限计算平均工资，同时将个人身份模糊化。

使用以下方法将该数据模糊化：

- 截断电话号码，只留区号。
- 隐藏名称和员工编号。
- 使用位移法，改变薪资和工作年限。
- 保留 IP 地址的前缀。

这种方法可以保护个人的薪资隐私，因此您可以放心地在云中计算平均工资，而无需担心数据和结果被暴露。使用与薪资和工作年限相关的位移位值可以在安全区域中调整结果。只要位移值保持机密状态，真实的平均工资将受到保护。截断电话号码可防止人们将其作为准标识符，并结合相关的数据条目来揭露个人身份。

概念验证

我们执行的概念验证（PoC）表明数据匿名化是一项切实可行的技术，可用于保护云计算。在概念验证（PoC）中，我们利用数据匿名化保护和处理存储在公共云中的事件日志数据。

为了能够更有效地与外部公司开展合作以及对我们的计算环境有更有认知，我们需要记录更多的数据日志。未来，其中的很多日志数据将在英特尔网络边界外或网络隔离区（demilitarized zone，简称 DMZ）生成。我们需要具备存储和分析这些日志数据的能力，以便日后使用。我们希望确定是否能够在保持数据安全之际，同时在公共云中使用软件即服务（software-as-a-service，简称 SaaS）日志管理供应商提供的应用。使用 SaaS 日志管理供应商提供的应用（而非简单地将日志数据存储在企业服务器上），让我们实现了快速、可检索和易于使用的日志归档，这将减少其它流程所需的日志条目，而且它采用了 Hadoop* 就绪格式，支持对海量数据的挖掘。[3]

PoC 实施

我们执行的概念验证（PoC）（花费了数周时间）的总体结构如下：

- 为存储匿名化数据评估匿名化工具。
- 在基于公共云的 SaaS 日志管理应用中存储匿名化数据。
- 使用 Hadoop 分析存储日志中的日志以收集安全和性能数据。
- 记录结果并报告。

我们使用现有的外部应用生成安全和性能数据，然后将其发送至 SaaS 日志管理供应商让其分析。我们选择的应用运行在学术云上，并监控大量网站的性能。从每台虚拟机（virtual machine，简称 VM）运行的网站上可以获得多项性能数据，例如：查看 web 服务器域名花费的时间，设置 TCP 连接花费的时间以及网页下载速度。图 2 说明了 SaaS 日志管理供应商、VM 和

[3] Hadoop* 项目为实现可靠、可扩展的分布式计算和数据存储提供了开源框架实施。

表 5. 示例：平均薪资计算和身份模糊化

IP 地址	名	姓	员工编号	工作年限	电话号码	月薪
143.183.79.169	x	y	0	11	408	15,000 美元
143.183.3.25	x	y	0	1	408	14,000 美元

存储数据匿名化转换表的安全区域之间的关系。

匿名化发生在发送数据的虚拟机上，在安全区域内数据将不再匿名化。在概念验证（PoC）过程中，47 台虚拟机以每小时 2,800 个事件（个人日志条目）的速度向 SaaS 日志管理供应商发送匿名日志信息。

我们决定匿名化 web 服务器访问日志中的 IP 地址和 web 服务器访问日志文件中的网址。

- 我们使用 IP 匿名的 perl 库隐藏 IP 地址。
- 我们使用 MD5 散列函数匿名化网址。虽然大多数安全行业认为 MD5 散列函数功能太弱而无法有效地发挥作用，但是我们仍然选择它作为易于实施的示例。在实际实施过程中将使用其它散列函数。
- 我们用于隐藏 IP 地址的匿名化软件采用了高级加密标准（Advanced Encryption Standard，简称 AES）加密以搞乱 IP 地址。

信息安全方面的分析

我们希望确定数据分发 web 服务器是否正在被探测。我们计划探测 web 服务器上的一些端口，然后查看能否检测到匿名化日志中的活动。检测安全异常的一种方法是检查日志活动是否出现激增。因为此种检测方法对于缓慢攻击来说并没有作用，所以我们又执行了一次探测，只对其中一个数据分发 web 服务器上的一个端口探测。

性能分析

我们还希望使用 SaaS 日志管理供应商提供的应用来研究我们正在分析的网站性能。举例来说，我们希望计算摘要统计数据，例如平均的 TCP 连接建立时间和这些时间之间的标准偏差。

成效

我们发现可以从在云中存储和分析的匿名化数据中收集有用的安全和性能信息。我们的模拟攻击在匿名化日志中可以被检测到，而且我们可以从匿名化的性能日志中发现性能问题。

- 在安全分析测试过程中，我们并未检测到对监控虚拟机的任何主动探测。但是，我们从搜索到的旧日志中发现 web 服务器上有探测器。这证实了我们的理论，即我们用于寻找安全商业智能事件的方法可以检测到真实事件。
- 尽管在概念验证（PoC）过程中使用的 SaaS 日志管理供应商应用并不支持计算平均值等数字运算分析，但是我们可以使用它来找出其他性能问题。例如，我们发现某个网站因为在用户访问实际内容之前执行了两次重定向，从而增加了访问时间。

主要经验

尽管在云计算中使用的匿名化方法在其投产之前还需要进行更多的研究，我们的概念验证（PoC）已证明其有用之处。下面介绍我们从概念验证（PoC）中总结的几点重要经验：

- 匿名化日志可以用于性能分析和安全分析。
- 匿名化可以缓解一些安全问题，简化隔离区（DMZ）和安全配置，并且支持在公共云中执行大量操作。
- 使用加密匿名化数据时，切记要管理好控制实际值到匿名值映射的加密密钥。

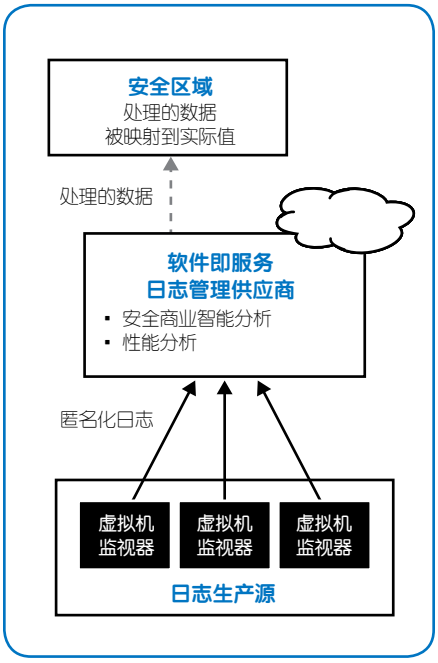

图 2. 我们的概念验证分析了云中的匿名化日志数据。

- 在云计算中的匿名化可以算得上是英特尔® 高级加密标准 — 新指令（Intel® Advanced Encryption Standard – New Instructions，简称 AES-NI）的一个新的重要用例，因为这项技术加快了某些匿名化技术（例如散列）使用的 AES 加密。

下一步工作

英特尔的 IT 安全架构和云工程组认为我们的概念验证（PoC）非常有用，足以支持我们对外部日志和数据匿名化开展更深入的研究。我们计划进一步完善匿名化的使用案例，利用英特尔数据再执行一次概念验证（PoC）。随着与我们合作的公司的日益增多以及对混合云愿景的不懈追求，我们的匿名化方法将有望成为帮助企业使用公共云的一种非常有用的技术。

今天，已经有利用匿名化技术的产品上市。这些产品可使公司不暴露机密信息的情况下使用公开托管的 SaaS 产品。但是，在匿名化方面，显然我们还需要做更多的工作。

- 我们需要让企业云用户了解匿名化及其潜在的优势和劣势。虽然数据匿名化并非万无一失，但是它适用于某些使用情况，而且我们有办法确定匿名化数据中是否存在漏洞。
- 市场上的开源工具功能强大但文档支持不足，一些已经被它们的创造者抛弃。我们打算为其中一些开源工具编制支持文档。
- 我们的概念验证（PoC）使用了 AES 加密来执行匿名化。英特尔 AES-NI 指令可用于加快匿名化过程。因此，公共云的安全使用有可能成为英特尔 AES-NI 的一个使用案例。我们将调整开源匿名化工具以使用英特尔 AES-NI，并调查它是否在匿名化过程中增添了价值。

总结

借助于企业私有云，英特尔实现了巨额的节省和成功。目前，我们已向混合云使用模式迈进，并积极采用多种技术来保护英特尔员工和客户的隐私。但是，若想使该模式有效运作，我们需要根据英特尔信息安全策略来保护数据。我们执行的概念验证（PoC）表明，数据匿名化 — 将已发布的数据弄得模糊来防止识别关键信息 — 是一款可提高云计算安全的可行技术。

为了能够更有效地与外部公司开展合作以及对我们的计算环境具有更多认知，英特尔 IT 部门需要记录更多的数据日志。未来，其中的很多日志数据将在英特尔网络边界外或隔离区（DMZ）生成。考虑到云基础设施并非完全安全，我们采用了数据匿名化方法积极应对云中的潜在安全漏洞，从而使数据对于其他人没有作用，但仍然允许英特尔 IT 部门有效地处理。

虽然数据匿名化并不是万无一失，但是对于不懈追求云计算安全的我们来说是一项非常重要的工具。我们打算进一步研究数据虚拟化，其中包括开发更多使用案例、让潜在的企业云用户了解数据匿名化的潜在优势和劣势、为现有开源数据匿名化应用编制支持文档，以及执行更多更广泛的概念验证（PoC）。

如欲了解有关英特尔 IT 部门最佳实践的更多信息，请访问：
www.intel.com/cn/it

缩写词

英特尔® AES-NI	英特尔® 高级加密标准 — 新指令
DMZ	隔离区
PoC	概念验证
SaaS	软件即服务
VM	虚拟机

0612/ACHA/KC/PDF 327464-001CN