大数据知识科普：中国数据量有多大？

这是一个注定要被拍砖的问题，但是这样拍着改着也可能就真明确了。所以无论如何，还是先抛出块砖头吧。

我们都在说大数据时代来临，信息和数据大爆炸。从 2013 年初开始，对于大数据爆发的焦虑感，紧迫感，不由自主地被卷入的甚至无力的感觉，驱动众多行业、企业和团体去关注和开始接触和了解大数据，自觉或不自觉的，主动或不得已地去融入这波洪流。但是，真的说到大数据，我们身边到底有多少数据量，它们都分布在哪些行业，哪些数据是目前可用的，哪些行业已经在使用数据，进入产业互联网和数据引导的变革了？

可能看到的版图依旧模糊。因此，我们怀抱很好的希望，以第一个吃螃蟹并期待来自行业的矫正和拍砖的态度，首先尝试对于国内各个领域，行业以及机构的数据拥有情况，使用情况以及未来路径做一个粗犷地调研、梳理和判断，对大数据时代我国各个领域数据资产的拥有和使用情况，也就是我们数据资产的家底做个盘点，也对各个行业、系统进军大数据，以及拥抱产业互联网的进度和未来做个简单判断。事实上，大数据之题无疑繁若星辰，然而只有在相对完整的视图下，繁星若尘，我们才可得以一窥天机。

闲话少叙，现在开始算账。

从我们手头掌握的数据来看，2013 年度，中国存储市场出货容量超过 1 个 EB，存储总量而 IDC 曾经发布的预测表明在未来的 3-4 年，中国存储总容量可能达到 18 个 EB。从数据存储市场的需求来看，互联网、医疗健康、通信、公共安全以及军工等行业的需求是主要的，且上升态势明显。

鉴于存储和服务器的紧密相关，我们从已经获得的资料可以知道，**目前全球运行的服务器总量超过 5000 万台，美国国内运行的服务器总体容量接近 1000 万台。**从各种市场公开数据来看，2013 年中国内地服务器销售总数接近为 100

万台。大体估算，**截止到 2013 年底，中国内地整体在运行的服务器总数量在 300 万台以上。**

从现有存储容量看，中国目前可存储数据容量大约在 8EB-10EB 左右，现有的可以保存下来的数据容量大约在 5EB 左右，且每两年左右会翻上一倍。这些被存储数据的大体分布为：媒体/互联网占据现有容量的 1/3，政府部门/电信企业占据 1/3，其他的金融、教育、制造、服务业各部分占据剩余 1/3 数据量。

公开数据显示，互联网搜索巨头百度 2013 年拥有数据量接近 EB 级别、阿里、腾讯声明自己存储的数据总量都达到了百 PB 以上。此外，电信、医疗、金融、公共安全、交通、气象等各个方面保存的数据量也都达到数十或者上百 PB 级别。

在目前被广泛引用的 IDC 和 EMC 联合发布的“2020 年的数字宇宙”报告预测到 2020 年，全球数字宇宙将会膨胀到 40000EB，均摊每个人身上是 5200GB 以上，这个量将会如何被有效存储和应用，我们眼下还很难想象。然而我们看到该报告指出，从现在起到 2020 年，全球数字宇宙的膨胀率大约为每两年翻一番。事实上，根据上述调查结论和服务器容量调查，我们也能做出个相对合理的推断：目前，全球产生的数据量中仅有 1%左右的数据能够被保存下来，也就是说今天全球能够被保存下来的数据也就是在 50EB 左右，而其中被标记并用于分析的数据更是不到 10%。

作为全球人口和计算设备保有量的大国，我国每年所能产生的数据量也极为庞大，有数据说 2014 年甚至可能达到 ZB 级别，但是真正被有效存储下来的数据仅仅是其中极微少部分，中国保存下来数据占全球数据的比例大约在 10%左右，也就是上面说的 5EB。这些数据中，目前已被标记并用于分析的数据仅达到 500PB 左右，也是接近 10%的一个比例。

伴随着云计算迅速普及和各行业，各企业和部门对于数据资产保存和利用意识的增强，以及通过互联网、大数据对产业进行变革的意愿，未来 2-3 年一定会有越来越多的行业、大企业步入到 PB、百 PB、甚至 EB 级别数据俱乐部，未来 3-3 年中国的数据总量也将呈翻倍上升态势，**我们预测 2015 年中国就可能突破**

10EB 数据保有量，被标签和分析利用数据量也将上升到 EB 级别，这些数据增长中互联网、政务、医疗、教育、安全等行业和领域所做贡献最大，而相对传统的物流、生产制造、甚至农业等领域数据拥有量的增长将更加明显。