ビッグデータ

ビッグデータは巨大なデータ?

ビッグデータ(Big Data)は、最近よく聞くバズワードである。バズワードとは、皆が使っていて、なんだかすごそうだけれど、実はあいまいでよくわかっていない流行言葉のことである。

バズワードではあるけれど、ビッグデータに関しては名前が示すように、従来の手法やツールで処理することが困難なほど巨大で複雑なデータの集合を指すと一般には理解されている。さらに、このようなデータをビジネスに役立てるために収集して分析することまでを含めて、単にビッグデータと呼ぶことも多い。

ちなみに、米国の調査大手のガートナー社は、以下に示す3つの特徴でビッグデータを定義している。

◯Volume:データ量の大きさ
◯Variety:データの種類の多様性
◯Velocity:データの発生頻度

いずれも英語で表記すると頭文字がVであることから、ビッグデータの特徴の3Vと呼ばれている。ちなみに、上記の3つに加えてさらに、

◯Veracity:データの正確性、データの取り方や発信媒体の信頼性なども含めた真実を反映している度合い

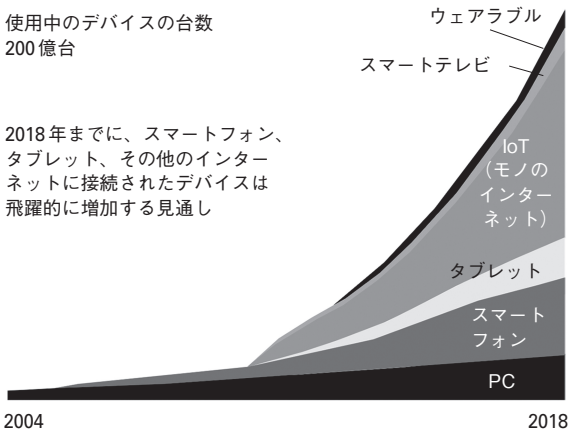
の4つを用いて4Vとする場合もある。

ビッグデータが広まった背景

ビッグデータが広まった理由の1つは、ICT(情報通信技術)の発達により、大量のデータを収集することが容易になったからである(**図1**)。

いまや、小売販売店のレジには必ず常備されているPOS(Point Of Sales:販売時点情報管理)システムや、別項のIoT(モノのインターネット)ではセンサ技術の発展および小型化・高性能化したセンサ機器によって、リアルタイムな情報収集の実現が容易になった。インターネットに接続可能な端末の増加を予測した**図1**に示す。急激に増加す

図1　インターネットに接続されたデバイスの爆発的な増加



[マルコ イアンシティ、カリム R. ラカニー:GEが目指すインダストリアル・インターネット、ダイヤモンド・ハーバード・ビジネス・レビュー(ダイヤモンド社)、2015年4月号、p.76より一部変更して引用]

表1　データ量を表す接頭語

キロ	Kilo	10^3	一千	400字詰め原稿1枚が約0.8キロバイト
メガ	Mega	10^6	百万	新聞の朝刊40ページすべて文字で埋めた場合が約0.8メガバイト
ギガ	Giga	10^9	十億	ヒトゲノム(全遺伝情報)の情報量が約0.8ギガバイト 長編映画1編をデジタル化して圧縮すると約1ギガバイト
テラ	Tera	10^{12}	一兆	ヒト一人の脳の容量が約1.2テラバイト
ペタ	Peta	10^{15}	千兆	世界最大の米議会図書館の全所蔵品(図書以外も含む)をデジタル化すると約0.3ペタバイト
エクサ	Exa	10^{18}	百京	世界にあるすべての印刷物の全情報量が約0.2エクサバイト 人類が過去に話したすべての言葉の情報量が約5エクサバイト
ゼタ	Zetta	10^{21}	十垓	世界に保存されている利用可能な全情報量が約1.2ゼタバイト (世界の砂丘の砂粒の数が約1ゼタ個)

る様子が理解できる。

次に、データ記録装置の低価格化によって、大量に発生するデータを低コストで蓄積することが可能になった。データ記録装置の代表であるHDD(ハードディスクドライブ)の性能向上を調査した資料によると、1981年には100万ドルだったHDDのギガバイト容量当たりの価格は、30年後の2011年には0.05ドルと激減した。この価格の低下は、30年間でほぼ一定のペースを守って推移している。計算すると、同容量のHDDの価格は毎年ほぼ約43%ずつ低下、つまり約4年で元の10%まで安価になる。

さらに、データ解析のアルゴリズムの高速化も大幅に進んだ。ある研究者の説によると、1983年から2003年までの20年間において、コンピュータのアルゴリズムは43,000倍もの高速化を実現したとの報告もある。

また、クラウド・コンピューティングと呼ばれている、インターネットなどのネットワークを通じたサービス提供を必要に応じて利用する方式の台頭によって、必ずしも自前でデータ蓄積や処理環境を用意する必要がなくなったこともビッグデータの広まりを後押ししている。

ビッグデータの量と単位

日本語の一文字分のデータ量、すなわち全角の漢字またはひらがなは2バイトである。これをもとに、ビッグデータの量を表す接頭語と、その量の目安を**表1**に示す。

ビッグデータが注目されたのは、インターネット上に作成されて蓄積された膨大なデータである。2002年に全世界で作成されたデータ量は23エクサバイトと推定されているが、わずか7年後の2009年には800エクサバイトまで急増した。今年2015年には8ゼタバイトと試算されており、2020年には45ゼタバイトにまで達するとの予想もある。増えたデータの大半は画像データやセンサーデータと言われているが、その75%は個人からのデータである。

国内でインターネット上に流通する情報量は、2012年時点で1カ月当たり2,300ペタバイトと推定されており、これはDVDメディア約5億7,000万枚分に相当する。

インターネット上の個人が作成する情報データ量は、SNS(Social Networking Service：インターネット上の交流を通して社会的なつながりをつくるサービス)の寄与も大きい。たとえば、SNSで毎日生み出されるデータ量は以下のように膨大である。

ツイッターでは、毎日4億件つまり毎秒4,600件を超えるつぶやき(ツイート)が行われている。ツイッターでのつぶやきの入力文字の上限は140文字であり、アルファベットに換算すると140バイトに相当するから、毎日約56ギガバイトの文字が生み出されて読まれていることになる。

フェイスブックでは、毎日の投稿数と「いいね！」クリック数の合計の平均は32億回、すなわち毎秒3万7,000回にも上る。なお、両社の数字は、いずれも2012年に公表されたものであり、現在はこれよりも増えていると推察される。