Tổng luận số 2 /2015

DỮ LIỆU LỚN VÀ XU HƯỚNG ĐỔI MỚI SÁNG TẠO DỰA TRÊN DỮ LIỆU

CỤC THÔNG TIN KHOA HỌC VÀ CÔNG NGHỆ QUỐC GIA
Địa chỉ: 24, Lý Thường Kiệt, Hoàn Kiếm, Hà Nội. Tel: (04)38262718, Fax: (04)39349127
Ban biên tập: TS. **Lê Xuân Định** (*Trưởng ban*), KS. Nguyễn Mạnh Quân,
ThS. Đặng Bảo Hà, ThS. Phùng Anh Tiến.

Mục lục

Lời giới thiệu

Thế giới đang chứng kiến một cuộc cách mạng công nghiệp mới được thúc đẩy bởi các dữ liệu số, tính toán và tự động hóa. Sự giao thoa của một số xu hướng công nghệ và kinh tế xã hội, bao gồm cả việc sử dụng Internet ngày càng tăng và sự suy giảm ở chi phí thu thập, truyền tải, lưu trữ và phân tích dữ liệu, dẫn đến việc tạo ra những khối lượng dữ liệu khổng lồ - gọi chung là "dữ liệu lớn" (Big Data), đây chính là nguồn lực có thể khai thác để thúc đẩy hình thành các ngành công nghiệp mới, các quy trình và sản phẩm mới. Các hoạt động kinh tế và xã hội từ lâu đã dựa vào dữ liệu. Tuy nhiên giờ đây, khối lượng, tốc độ và chủng loại dữ liệu được sử dụng đang gia tăng mạnh mẽ trên phạm vi toàn bộ nền kinh tế, và quan trọng hơn là giá trị kinh tế và xã hội lớn hơn của chúng đang mở ra cơ hội về một sự thay đổi hướng tới mô hình kinh tế xã hội dựa trên dữ liệu. Trong mô hình này, dữ liệu là tài sản cốt lõi có thể tạo ra lợi thế cạnh tranh quan trọng, chi phối đổi mới sáng tạo, tăng trưởng và phát triển bền vững.

Đổi mới sáng tạo dựa vào dữ liệu có giá trị kinh tế to lớn, với doanh thu từ các sản phẩm và dịch vụ Dữ liệu lớn đã vượt quá 18 tỷ USD trong năm 2013, và theo Feff Kelly (2014) thì giá trị này có thể đạt 50 tỷ USD vào năm 2017. Để hiện thực hóa trọn vẹn tiềm năng của dữ liệu lớn, các quốc gia cần có một khuôn khổ chính sách nhất quán, phù hợp về thu thập, truyền tải, lưu trữ, cung cấp và sử dụng dữ liệu, đặc biệt là trong các lĩnh vực như bảo vệ quyền riêng tư, tiếp cận dữ liệu mở, kỹ năng và việc làm, cơ sở hạ tầng và đo lường, v.v... Đây cũng chính là những nội dung thông tin mà cuốn Tổng luận **"Dữ liệu lớn và xu hướng đổi mới sáng tạo dựa trên dữ liệu"** muốn cung cấp với các độc giả. Tài liệu được biên soạn dựa trên các báo cáo của OECD về vai trò tiềm năng của dữ liệu và phân tích dữ liệu trong việc tạo ưu thế cạnh tranh và hình thành vốn tri thức, thúc đẩy đổi mới sáng tạo và tăng trưởng bền vững. Phần đầu của Tài liệu cung cấp những khái niệm và định nghĩa đã được công nhận rộng rãi về Dữ liệu lớn, cũng như việc tạo ra và sử dụng dữ liệu trong các lĩnh vực ứng dụng của nền kinh tế. Tiếp theo tài liệu mô tả các cách thức khai thác dữ liệu như một nguồn lực thúc đẩy tăng trưởng kinh tế và phát triển bền vững, và trong phần cuối, tài liệu đề cập đến các vấn đề chính sách chủ yếu trong hoạch định chính sách công nhằm thúc đẩy đổi mới sáng tạo dựa vào dữ liệu.

Xin trân trọng giới thiệu.

CỤC THÔNG TIN KH&CN QUỐC GIA

Bảng các chữ viết tắt

API	Giao diện lập trình ứng dụng
BI	Trí tuệ doanh nghiệp
CAGR	Tỷ lệ tăng trưởng tổng hợp lũy kế hàng năm
DDI	Đổi mới sáng tạo dựa vào tăng trưởng
HDD	Ổ đĩa cứng
ICT	Công nghệ thông tin - truyền thông
IoT	Internet kết nối vạn vật
KBC	Vốn tri thức
M&A	Mua bán và sáp nhập
M2M	Giao tiếp máy tới máy
NC&PT	Nghiên cứu và phát triển
NoSQL	Cơ sở dữ liệu phân tán không quan hệ
OECD	Tổ chức hợp tác và phát triển kinh tế
PET	Công nghệ bảo vệ quyền riêng tư
PMNM	Ứng dụng phần mềm nguồn mở
PSI	Thông tin khu vực công
SHTT	Sở hữu trí tuệ
SMS	Tin nhắn văn bản
SSD	Ổ đĩa thể rắn

I. ĐỔI MỚI DỰA TRÊN DỮ LIỆU - NGUỒN LỰC TĂNG TRƯỞNG VÀ PHÁT TRIỂN KINH TẾ

1.1. Dữ liệu lớn và các khái niệm liên quan

Trong thời đại hiện nay, dữ liệu đang ngày càng thấm sâu vào cuộc sống của con người hơn bao giờ hết. Chúng ta mong muốn sử dụng dữ liệu để giải quyết các vấn đề, nâng cao phúc lợi và tạo ra thịnh vượng kinh tế. Việc thu thập, lưu trữ, và phân tích dữ liệu đang tuân theo quỹ đạo có xu hướng đi lên và dường như không có ranh giới, hoạt động này được thúc đẩy bằng những gia tăng về năng lực xử lý, chi phí giảm mạnh trong tính toán và lưu trữ, và số lượng ngày càng tăng các công nghệ cảm biến nhúng trong tất cả các loại thiết bị. Vào năm 2011, một số ước tính rằng khối lượng thông tin được tạo ra và sao chép lại sẽ vượt mức 1,8 zettabytes. Trong năm 2013, ước tính có 4 zettabytes dữ liệu được tạo ra trên toàn thế giới.

1 zettabyte (ZB) = 10^{21}bytes. Một byte tương đương với một ký tự trong văn bản. Có thể tưởng tượng rằng, nếu cứ mỗi giây, mỗi một người dân tại Hoa Kỳ chụp một bức ảnh số, cứ thế liên tục trong vòng một tháng. Tất cả số ảnh đó đem tập hợp lại với nhau sẽ bằng khoảng một zettabyte.

Mỗi ngày có hơn 500 triệu bức ảnh được tải lên và chia sẻ trên mạng xã hội, cùng với các đoạn video với độ dài tổng cộng đến 200 giờ được tải lên mỗi phút. Nhưng khối lượng thông tin mà mọi người tự tạo ra, các thông tin liên lạc gồm các cuộc gọi thoại, email và văn bản, các bức ảnh, video và âm nhạc được tải lên vẫn không là gì so với lượng thông tin số được tạo ra về chúng mỗi ngày.

Các xu hướng này vẫn đang tiếp diễn. Hiện nay chúng ta mới ở vào giai đoạn rất sơ khai của cái gọi là "Internet vạn vật" (IoT), khi tất cả các thiết bị, các phương tiện và các công nghệ "mang trên người" có thể giao tiếp được với nhau. Các tiến bộ công nghệ sẽ làm giảm chi phí của việc tạo ra, thu thập, quản lý và lưu trữ thông tin xuống chỉ còn bằng một phần sáu chi phí được tính vào năm 2005. Và kể từ năm 2005, đầu tư doanh nghiệp vào phần cứng, phần mềm, nhân lực và dịch vụ đã tăng 50% đạt 4 nghìn tỷ USD.

"Internet vạn vật" là thuật ngữ dùng để mô tả khả năng các thiết bị có thể giao tiếp được với nhau sử dụng các cảm biến nhúng, liên kết với nhau thông qua các mạng kết nối có dây và không dây. Các thiết bị này có thể bao gồm cả nhiệt kế, xe hơi và thậm chí cả viên thuốc mà bạn nuốt vào để các bác sĩ có thể theo dõi sức khỏe bộ máy tiêu hóa của bạn. Các thiết bị kết nối này sử dụng Internet để truyền, diễn giải và phân tích dữ liệu.

1.1.1. Dữ liệu và các yếu tố thúc đẩy tạo và sử dụng dữ liệu

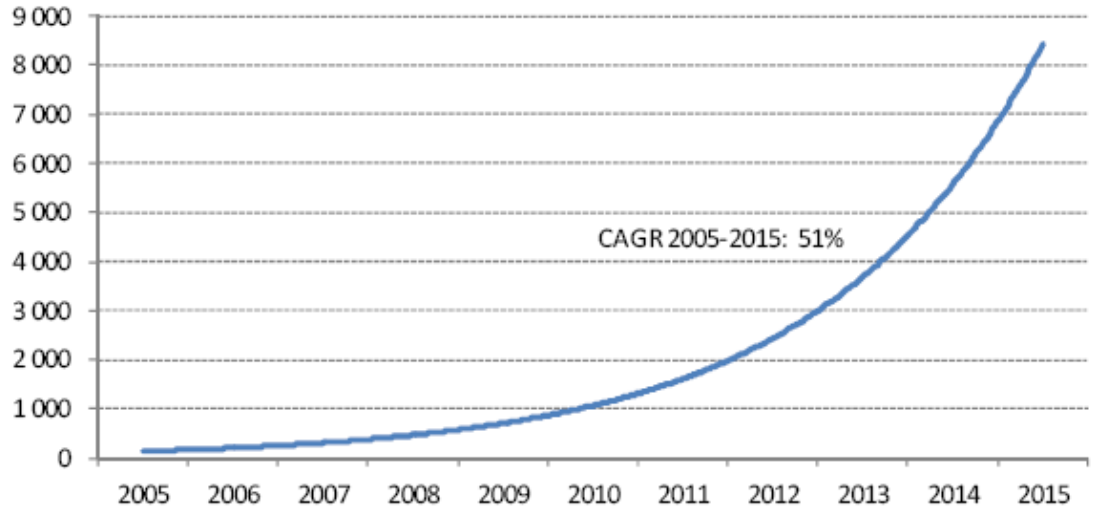
Việc số hóa gần như mọi phương tiện truyền thông và sự chuyển hướng ngày càng tăng

của các hoạt động kinh tế và xã hội sang sử dụng Internet (thông qua các dịch vụ điện tử như các mạng xã hội, thương mại điện tử, y tế điện tử và chính phủ điện tử) đang tạo ra nhiều petabyte (hàng triệu gigabyte) dữ liệu cứ sau mỗi giây. Ví dụ như mạng kết nối xã hội Facebook được biết có đến 900 triệu người tham gia trên toàn thế giới và tạo ra trung bình hơn 1500 trạng thái cập nhật mỗi giây (Hachman, 2012; Bullas, 2011).

Với việc khai thác và kết nối (thế giới thực) ngày càng tăng của các bộ cảm biến thông qua các mạng cố định và di động (mạng cảm biến), ngày càng có nhiều các hoạt động ngoại tuyến cũng được ghi lại bằng kỹ thuật số, dẫn đến một làn sóng bổ sung dữ liệu không ngừng.

Nhiều tài liệu chỉ ra rằng, riêng trong năm 2010, các doanh nghiệp lưu trữ tổng thể hơn 7 exabyte (hàng tỷ gigabyte) dữ liệu mới trên các ổ đĩa, trong khi người tiêu dùng bảo quản hơn 6 exabyte dữ liệu mới (MGI, 2011). Điều đó dẫn đến một lượng dữ liệu tích lũy ước tính hơn 1000 exabyte vào năm 2010; một nhà phân tích ước tính rằng con số này sẽ tăng lên gấp 40 lần vào cuối thập kỷ này (IDC, 2012).

Hình 1: Kho dữ liệu ước tính trên phạm vi toàn thế giới, đơn vị exabyte (tỷ gigabyte)



Nguồn: OECD dựa trên dự báo nghiên cứu của IDC Digital Universe.

Tạo dữ liệu, thu thập và truyền tải

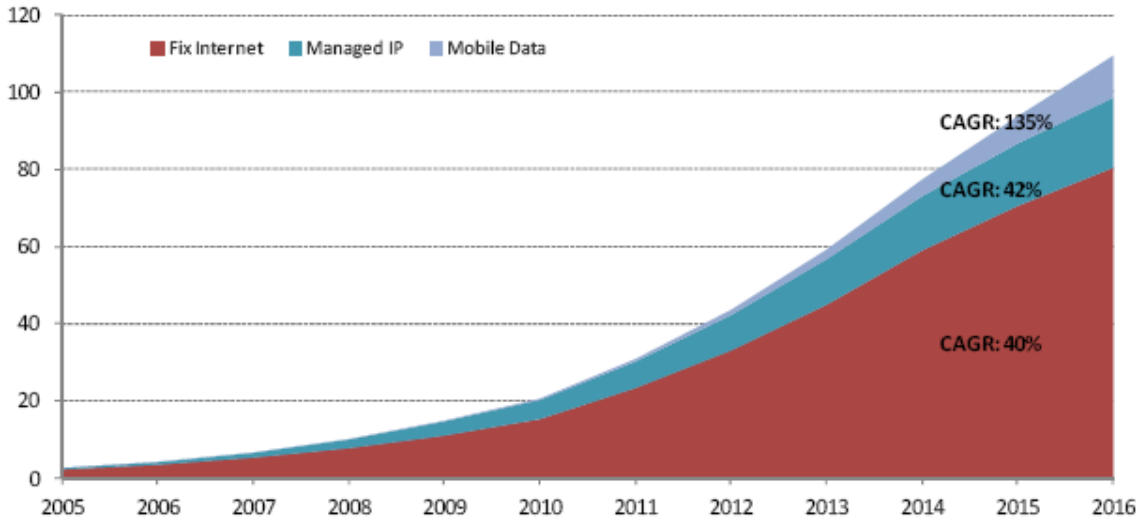
Lượng dữ liệu gia tăng một cách mạnh mẽ chủ yếu bị tác động bởi sự hội tụ của những phát triển công nghệ quan trọng, đáng chú ý là truy cập băng thông rộng ở mọi nơi và sự phổ biến các thiết bị và ứng dụng ICT thông minh, như các dụng cụ đo thông minh, lưới điện và giao thông vận tải thông minh dựa trên các mạng cảm biến và sự giao tiếp máy với máy (M2M). Chi phí truy cập Internet giảm mạnh trong vòng 20 năm qua là một yếu tố chi phối quan trọng. Ví dụ vào năm 2011, người tiêu dùng ở Pháp phải trả khoảng 33 USD một tháng cho một kết nối băng thông rộng tốc độ 51 Mbit/s, trong khi chi phí cho

một kết nối bằng quay số (với tốc độ chậm hơn đến 1000 lần) là 75 USD vào năm 1995. Điện thoại di động đã trở thành một thiết bị thu thập dữ liệu hàng đầu, kết hợp dữ liệu định vị địa lý với kết nối Internet để hỗ trợ các dịch vụ trên phạm vi rộng và ứng dụng mới liên quan đến giao thông, môi trường và y tế. Nhiều dịch vụ và ứng dụng đó dựa (hoặc tham gia vào) việc thu thập và sử dụng dữ liệu cá nhân. Bổ sung cho sự truy cập Internet ngày càng gia tăng và hiệu quả hơn, hầu hết các thiết bị di động được trang bị các mảng giao thức gia tăng để trao đổi dữ liệu cục bộ (như Wifi, Bluetooth, Near Field Communications (NFC) với khả năng truyền dữ liệu ngang hàng (peer-to-peer). Các thiết bị này còn có thể quay video, chụp ảnh và ghi âm thanh (thường gắn với thông tin định vị).

Vào năm 2011, toàn thế giới có gần sáu tỷ thuê bao di động, trong đó khoảng 13% (780 triệu) là điện thoại thông minh có khả năng thu thập và truyền dữ liệu định vị địa lý (ITU, 2012; Cisco, 2012). Cũng vào năm này, các thiết bị điện thoại di động tạo ra khoảng 600 petabyte (triệu gigabyte) dữ liệu mỗi tháng (Cisco, 2012). Với sự phổ cập điện thoại di động (số thuê bao trên 100 dân) vượt quá 100% tại hầu hết các nước OECD và sự phổ biến băng thông rộng không dây đạt gần 50%, thì nguồn dữ liệu này sẽ gia tăng đáng kể khi mà điện thoại thông minh trở thành thiết bị cá nhân phổ biến. Cisco (2012) ước tính rằng lưu lượng dữ liệu sản sinh ra từ điện thoại di động sẽ đạt gần 11 exabyte (hàng tỷ gigabyte) vào năm 2016, có nghĩa là tăng gần gấp đôi mỗi năm (xem hình 2).

Hình 2: Lưu lượng IP toàn cầu hàng tháng, 2005-16.

Đơn vị: exabyte (1 exabyte = 1 tỷ gigabyte)



Nguồn: OECD dựa trên số liệu của Cisco (2012).

Sự gia tăng dữ liệu di động không chỉ do sự gia tăng số điện thoại di động, được dự báo sẽ chiếm đến một nửa tổng lưu lượng di động vào năm 2016 (Cisco, 2012). Các thiết bị

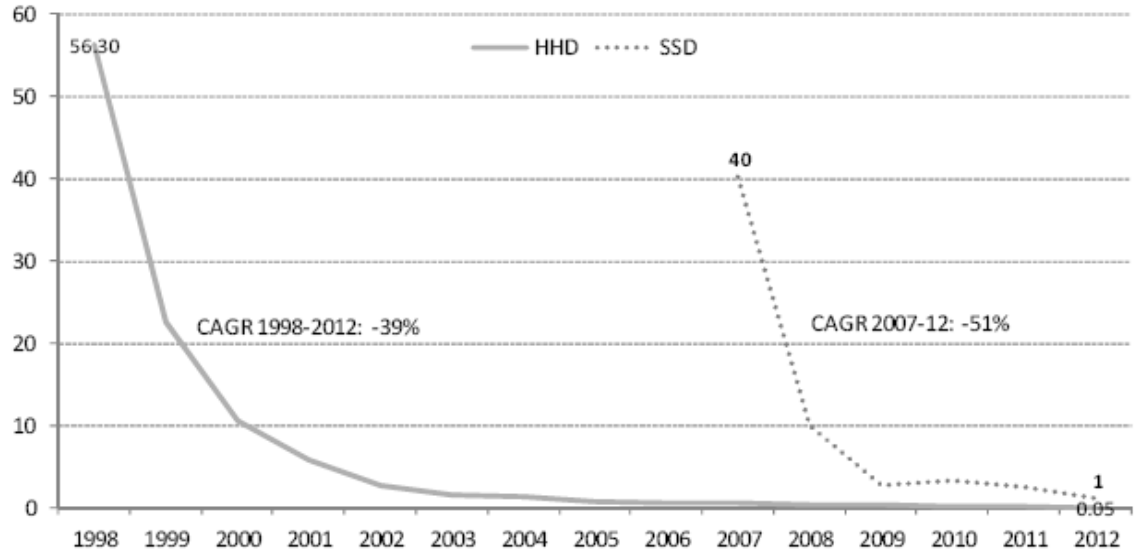
thông minh khác đang phát triển thậm chí còn nhanh hơn. Ví dụ, các dụng cụ đo thông minh thu thập và truyền dữ liệu thời gian thực ngày càng tăng (OECD, 2012), và xe ô tô thông minh giờ đây đã có thể truyền dữ liệu thời gian thực về hiện trạng các linh kiện trong xe và về môi trường (OECD, 2012). Nhiều thiết bị thông minh trong số này được dựa trên cơ sở các mạng kết nối cảm biến và thiết bị đi kèm có thể cảm nhận và tương tác với môi trường thông qua các mạng di động. Các bộ cảm biến và thiết bị đi kèm trao đổi dữ liệu thông qua các kết nối không dây "tạo khả năng tương tác giữa con người hay máy tính với môi trường xung quanh" (Verdone et al., 2008). Hơn 30 triệu bộ cảm biến kết nối tương tác hiện đang được triển khai trên phạm vi toàn thế giới trong các lĩnh vực như an ninh, y tế, môi trường, các hệ thống giao thông vận tải hay hệ thống kiểm soát năng lượng, số lượng của chúng đang tăng lên với tỷ lệ khoảng 30% một năm (MGI, 2011).

1.1.2. Lưu trữ và xử lý dữ liệu

Nếu như những phát triển công nghệ nêu trên chủ yếu thúc đẩy sự sản sinh và truyền tải dữ liệu, thì việc sử dụng dữ liệu đã trở nên dễ dàng hơn nhiều nhờ vào sự giảm mạnh chi phí lưu trữ, xử lý và phân tích dữ liệu. Trước đây, chi phí lưu trữ dữ liệu đã không khuyến khích việc giữ lại dữ liệu đã không còn hoặc có vẻ như không còn cần thiết (OECD, 2011). Nhưng chi phí lưu trữ đã giảm đến mức thấp để có thể lưu trữ dữ liệu trong thời gian dài, thậm chí là vô thời hạn. Điều này có thể được minh họa qua chi phí trung bình cho mỗi gigabyte ổ đĩa cứng (HDD), chi phí này đã giảm từ 56 USD năm 1998 xuống 0,05 USD năm 2012, tốc độ giảm trung bình hàng năm là gần 40% (xem hình 3). Với các công nghệ lưu trữ thế hệ mới như ổ đĩa thể rắn (SSD) chẳng hạn, chi phí trên mỗi gigabyte thậm chí còn giảm nhanh hơn.

Hình 3: Chi phí trung bình lưu trữ dữ liệu cho người tiêu dùng, 1998-2012

Đơn vị: USD/gigabyte

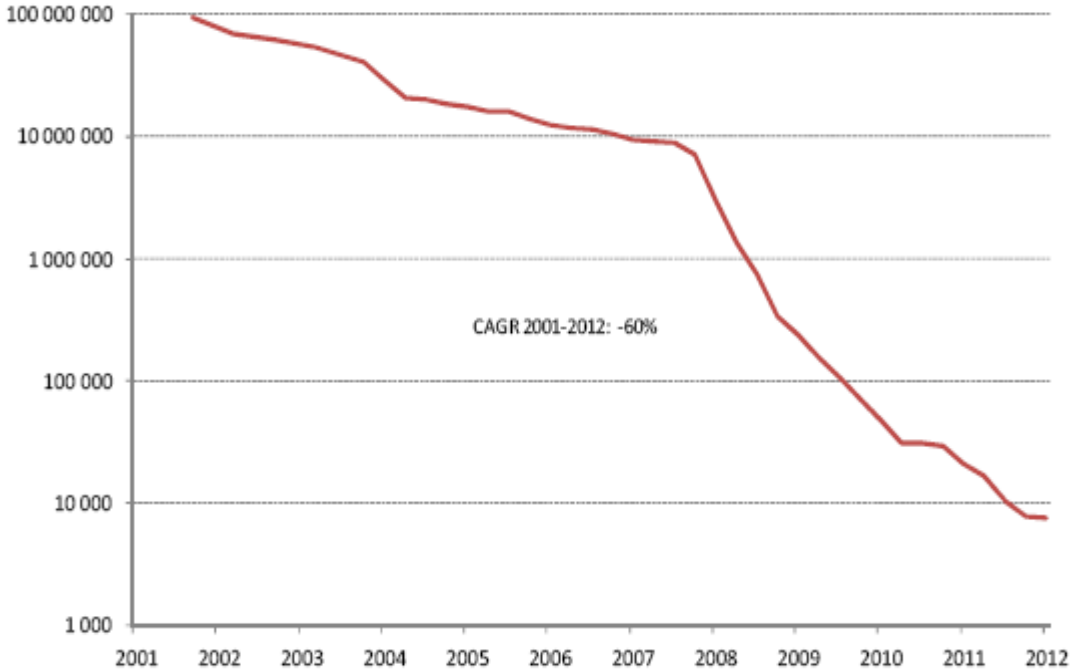


Nguồn: OECD trên cơ sở Pingdom (2011).

Định luật Moore phát biểu rằng tính năng xử lý tăng gấp đôi cứ sau 18 tháng, liên quan đến chi phí hay độ lớn chủ yếu đã được xác minh. Điều này đặc biệt đáng chú ý đối với các công cụ xử lý dữ liệu, chúng ngày càng trở nên có tính năng mạnh, tinh xảo, hiện diện mọi nơi và có giá rẻ, tạo điều kiện dễ dàng tìm kiếm dữ liệu, kết nối và truy xuất nguồn gốc, không chỉ các chính phủ và các tập đoàn lớn mà nhiều người khác đều có thể thực hiện được. Ví dụ như trong lĩnh vực di truyền, các máy lập trình tự gen ADN giờ đây có thể đọc được khoảng 26 triệu ký tự mã di truyền ở người trong chưa đầy một phút, và chi phí lập trình tự mỗi bộ gen đã giảm 60% một năm, trung bình từ 100 triệu USD năm 2001 xuống chưa đến 10.000 USD vào năm 2012 (xem hình 4).

Hình 4: Chi phí lập trình tự bộ gen, 2001-11

Đơn vị USD (theo thang đo logarit)



Nguồn: OECD dựa theo Viện nghiên cứu bộ gen người quốc gia Hoa Kỳ (www.genome.gov/sequencingcosts/)

Điện toán đám mây đóng vai trò quan trọng trong việc gia tăng khả năng lưu trữ và xử lý dữ liệu. Nó được mô tả như một "mô hình dịch vụ tính toán dựa trên một tập hợp tài nguyên máy tính có thể truy cập theo cách thức linh hoạt, mềm dẻo và theo nhu cầu với yêu cầu quản lý thấp" (OECD, 2012). Đặc biệt, đối với các doanh nghiệp vừa và nhỏ (SMEs), và cả các chính phủ không thể hoặc không muốn thực hiện những đầu tư lớn, phải thanh toán trước cho các công nghệ ICT, điện toán đám mây mang lại khả năng cho các tổ chức chi trả cho các nguồn lực siêu tính toán theo phương thức chi tiêu tùy theo khả năng (pay-as-you-go).

Các ứng dụng phần mềm nguồn mở (PMNM) bao gồm đầy đủ các giải pháp cần thiết cho dữ liệu lớn, chẳng hạn như để lưu trữ, xử lý và phân tích (bao gồm cả hiển thị trực quan - visualization), cũng góp phần đáng kể vào việc làm cho phân tích dữ liệu lớn có thể tiếp cận đến dân số rộng lớn hơn. Nhiều công cụ dữ liệu lớn được các công ty Internet phát triển ban đầu giờ đây được phổ biến rộng khắp nền kinh tế tạo ra các hàng hóa và dịch vụ mới dựa vào dữ liệu. Ví dụ, Hadoop, khung lập trình mã nguồn mở để quản trị dữ liệu phân tán, được lấy cảm hứng từ một bài báo của các nhân viên Google, Dean và Ghemawat (2004). Ban đầu nó được Yahoo! tài trợ và được các công ty Internet như Amazon, Facebook 11, 12 và LinkedIn khai thác và tiếp tục phát triển, sau đó được cung cấp bởi các nhà cung cấp cơ sở dữ liệu và máy chủ doanh nghiệp truyền thống như IBM, Oracle, Microsoft, và SAP như là một phần dòng sản phẩm của họ, và hiện đang được sử dụng rộng rãi cho các hoạt động dữ liệu chuyên sâu tại các doanh nghiệp thuộc đủ các loại như Wal-Mart (bán lẻ), Chevron (năng lượng) và Morgan Stanley (dịch vụ tài chính).

Ngày càng có nhiều nhà phân tích dữ liệu chuyên môn hóa và các nhà môi giới dữ liệu chào mời dữ liệu để sử dụng cho các mục đích như quảng cáo, kiểm tra lý lịch tuyển dụng việc làm, cấp tín dụng và thực thi pháp luật. Số các doanh nghiệp chào bán dữ liệu đã tăng mạnh trong những năm gần đây. Tại thời điểm năm 2013, tổ chức privacyrights.org đã liệt kê chỉ riêng ở Hoa Kỳ có đến 180 công ty môi giới dữ liệu trực tuyến đăng ký. Các hãng môi giới dữ liệu rất đa dạng, từ các công ty chuyên môn hóa giữa các doanh nghiệp (business-to-business) đến các dịch vụ nội bộ hóa đơn giản. Có thể kể đến các công ty như LexisNexis đã từng tuyến bố họ tiến hành hơn 12 triệu kiểm tra lý lịch một năm, và BlueKai Exchange tuyên bố là thị trường dữ liệu lớn nhất thế giới cho các nhà quảng cáo, công ty này sở hữu dữ liệu về hơn 300 triệu người tiêu dùng và hơn 30.000 thuộc tính dữ liệu. Theo thông tin công bố trên trang web của mình, BlueKai Exchange cho biết họ xử lý hơn 750 triệu sự kiện dữ liệu và giao dịch, thực hiện hơn 75 triệu cuộc bán đấu giá các thông tin cá nhân mỗi ngày.

1.1.3 Định nghĩa dữ liệu lớn

Có nhiều định nghĩa về "dữ liệu lớn" (Big data), và chúng có thể khác nhau tùy thuộc vào việc bạn là nhà khoa học máy tính, nhà phân tích tài chính hay một doanh nhân đang thuyết minh ý tưởng đầu tư mạo hiểm.

Nhiều tác giả mô tả đơn giản "dữ liệu lớn" như những kho chứa dữ liệu lớn (Large pools of data) (McGuire *et al*., 2012). Loukides (2010) định nghĩa đó là dữ liệu mà trong đó "*chính bản thân độ lớn của dữ liệu đã trở thành một phần của vấn đề*". Viện Nghiên cứu toàn cầu McKinsey (*McKinsey Global Institute* - MGI) cũng đưa ra định nghĩa tương tự "*đó là dữ liệu có độ lớn vượt quá khả năng các công cụ phần mềm cơ sở dữ liệu tiêu biểu có thể nắm bắt, lưu trữ, quản trị và phân tích*".

Hầu hết các định nghĩa phản ánh năng lực công nghệ ngày càng gia tăng để nắm bắt, tổng hợp và xử lý khối lượng dữ liệu với độ lớn, tốc độ và sự đa dạng lớn chưa từng thấy. Nói theo cách khác, "*dữ liệu giờ đây được cung cấp nhanh hơn, độ bao phủ và phạm vi*