DỮ LIỆU LỚN VÀ VAI TRÒ CỦA NÓ TRONG GIÁO DỤC ĐẠI HỌC

Vũ Hưng Hải*

Ngày nhận: 27/10/2014
Ngày nhận bản sửa: 6/11/2014
Ngày duyệt đăng: 25/11/2014

Tóm tắt:

Dữ liệu lớn (Big Data), một trong ba lĩnh vực của công nghệ thông tin được cho là đang và sẽ có ảnh hưởng khắp mọi nơi (hai lĩnh vực còn lại là thiết bị thông minh - Smart Device, điện toán đám mây - Cloud Computing). Theo một nghiên cứu của Nhật Bản về thị trường châu Á, các công ty có thể tăng thu nhập từ 25% tới 65% bằng việc dùng Big Data và họ cũng dự đoán thị trường tiềm năng của dữ liệu lớn có thể đạt tới 250 tỷ USD vào năm 2020 (Vu, 2014). Mọi khía cạnh trong đời sống chúng ta đều sẽ bị ảnh hưởng bởi dữ liệu lớn. Vậy dữ liệu lớn có vai trò gì trong lĩnh vực giáo dục đại học? Bài viết phân tích vai trò của dữ liệu lớn trong giáo dục đại học và đề xuất một số khuyến nghị nhằm ứng dụng và khai thác dữ liệu lớn trong các tổ chức giáo dục đại học ở Việt Nam.

Từ khóa: Dữ liệu lớn, Vai trò, Giáo dục đại học

Big Data and its role in Higher Education

Abstract:

Big Data, one of the three Information Technology fields, is believed to have had a worldwide impact (the other two fields are Smart Devices and Cloud Computing). According to a Japan's study on the Asian market, the use of Big Data may enable companies to increase their revenues from 25% to 65%, and they also predict the potential market for big data can reach 250 billion in 2020 (Vu, 2014). Every aspect of life will be affected by Big Data. So what role does Big Data play in higher education? This paper analyses the role of big data in higher education and recommends solutions for utilizing big data in Vietnam's higher education institutions.

Keywords: Big Data, Role, Higher Education

1. Giới thiệu

Dữ liệu lớn (*Big Data*) đề cập đến việc thu thập khối lượng dữ liệu khổng lồ mà một công ty/tổ chức có trong nội bộ, tổ hợp chúng với các nguồn dữ liệu bên ngoài như *Internet* rồi phân tích chúng để thu được thông tin có giá trị như *xu hướng* và *cơ hội*. Các công ty toàn cầu cần tới Big Data để nhận diện thị trường kinh doanh mới và dùng các phân tích để trợ giúp một cách hiệu quả cho nhà quản lý ra quyết định. Chẳng hạn, người phân tích Dữ liệu lớn (*Big Data Analytics*) dùng phân tích xu thế để hiểu chiều hướng thị trường, nhu cầu khách hàng để phát triển sản phẩm mới trước khi người khác thậm chí biết về nó. Dữ liệu lớn cũng giúp phân tích xu hướng của lĩnh vực và thị trường tương lai để quản lý tổng thể

chuỗi cung ứng: từ nguyên vật liệu thô tới chế tạo, từ các kênh phân phối cho tới các đại lý bán lẻ, cắt giảm chi phí vận hành và tối đa hóa lợi nhuận.

Trong bối cảnh suy thoái kinh tế toàn cầu ít người dám làm nước đi táo bạo nào, nhưng với việc dùng Big Data, một số công ty thương mại có khả năng nhận diện những xu hướng nào đó một cách nhanh chóng và nắm lấy cơ hội. Rất nhiều công ty đầu tư cho Big Data. Họ nhìn thấy việc làm chủ được Big Data sẽ cho phép giải quyết nhiều vấn đề phức tạp trước kia không thể làm được hoặc có thể tạo ra các quyết định và hành động tốt hơn. Điều này cho phép họ có được các ưu thế cạnh tranh, vấn đề cốt tử trong bối cảnh toàn cầu hóa. Hơn nữa, việc làm chủ dữ liệu lớn từ các mạng xã hội cho phép thấu hiểu các hành vi phức tạp của xã hội con người và nhiều hy vọng ở những đột phá trong khoa học công nghệ. Một nghiên cứu của Marr (2013) chỉ ra các ứng dụng dữ liệu lớn được sử dụng phổ biến nhất, tạo ra được những lợi ích cao nhất trong 10 lĩnh vực: *1) Sự hiểu biết và khách hàng mục tiêu 2) Sự hiểu biết và tối ưu hóa quy trình kinh doanh 3) Định lượng cá nhân và tối ưu hóa hiệu suất 4) Cải thiện chăm sóc sức khỏe và y tế công cộng 5) Cải thiện hiệu suất thể thao 6) Nâng cao khoa học và nghiên cứu 7) Tối ưu hóa hiệu suất máy móc thiết bị 8) Cải thiện an ninh và thực thi pháp luật 9) Cải thiện và tối ưu hóa các thành phố, quốc gia 10) Hỗ trợ các giao dịch tài chính, kinh doanh.*

Ngoài các lĩnh vực trên, các tổ chức giáo dục ngày càng quan tâm tận dụng lợi thế của dữ liệu lớn nhằm cải thiện kết quả học tập của sinh viên và nâng cao hiệu quả giảng dạy, nghiên cứu của giảng viên, trong khi giảm khối lượng công việc hành chính. Dữ liệu về sinh viên ngày càng được lưu giữ và quan tâm nhiều hơn. Những dữ liệu này có thể được bổ sung với dữ liệu hành vi thu thập từ các nguồn phương tiện truyền thông xã hội, các blog, các cuộc điều tra về sinh viên, các nguồn dữ liệu công cộng cũng như dữ liệu ở các tổ chức đại học. Với các nguồn dữ liệu trên, các tổ chức giáo dục có thể đánh giá sinh viên, giảng viên, giáo trình học liệu nhằm mang lại cái nhìn sâu sắc hơn, giúp cho việc cải thiện và nâng cao hiệu quả quản trị của họ (Schmarzo, 2014).

Với khối lượng lớn thông tin về sinh viên, bao gồm tuyển sinh, kết quả học tập và rèn luyện, các trường đại học có các bộ dữ liệu cần thiết để thu được lợi ích từ các dự án phân tích theo mục tiêu. Các báo cáo gần đây cho rằng các quản trị viên trên khắp nước Mỹ đã nhận ra tiềm năng của các phân tích và đã đầu tư nhiều nguồn lực hơn vào các dự án dữ liệu lớn. Một khảo sát của Gartner (2013) về "*Dữ liệu lớn, các cơ hội lớn hơn: Đầu tư vào Thông tin và Phân tích*" cho thấy 42% các chuyên gia công nghệ thông tin trên tất cả các lĩnh vực báo cáo đầu tư vào các dự án dữ liệu lớn hoặc có kế hoạch làm như vậy trong những năm tới. Báo cáo cũng xác định giáo dục đại học là lĩnh vực có triển vọng rất lớn cho các ứng dụng dữ liệu lớn, với lý do sự sẵn có của các nguồn dữ liệu khác nhau (Nagel, 2013).

Tác giả cho rằng, dữ liệu lớn cũng đóng vai trò lớn trong lĩnh vực giáo dục đào tạo ở Việt Nam, vì thế việc nghiên cứu để khai thác nó là cần thiết. Bài viết sẽ tìm hiểu về dữ liệu lớn, xu hướng, vai trò cũng như những ảnh hưởng của nó trong giáo dục đại học. Các phân tích và kết luận sẽ định hướng các cơ sở giáo dục, nhất là các trường đại học quan tâm nghiên cứu, khai thác dữ liệu lớn một cách hiệu quả phục vụ tốt hơn cho các mục tiêu đặt ra hiện nay và tương lai.

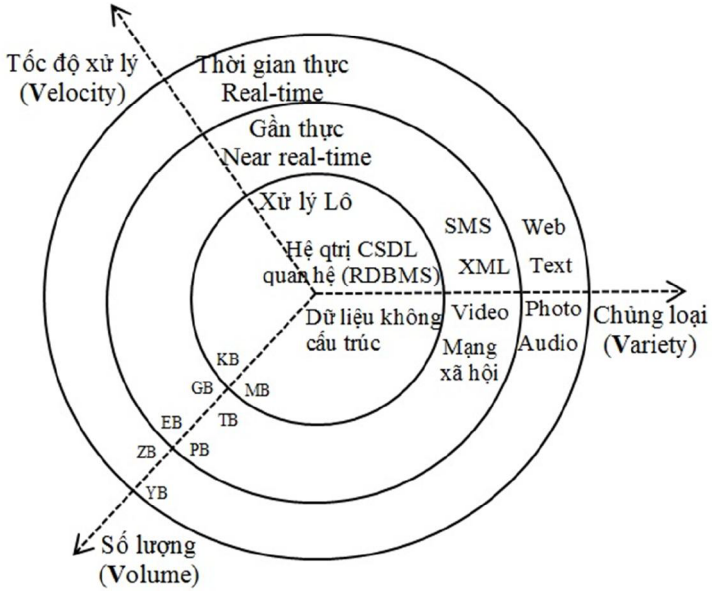
2. Tổng quan về dữ liệu lớn

2.1. Mô hình và tính chất

Thuật ngữ Big Data (*Dữ liệu lớn*) xuất hiện lần đầu trong bài báo của John R. Masey tại USENIX meeting (1998) với nhan đề "*Big Data... and the Next Wave of Infrastress*". Tháng 8/1999, Steve Bryson cùng đồng nghiệp có bài viết "*Visually exploring gigabyte data sets in real time*" trong hội thảo Communications of the ACM (CACM). Đây là lần đầu tiên CACM dùng thuật ngữ "Big Data"- tiêu đề của một trong các phần hội thảo là "*Big Data for Scientific Visualization*" (Press, 2013).

Big Data có nhiều định nghĩa. Đó *là thuật ngữ chỉ tập hợp dữ liệu lớn và phức tạp đến mức khó có thể xử lý bằng các công cụ hay ứng dụng quản trị và phân tích dữ liệu truyền thống* (White, 2012). Thách thức công nghệ nằm ở các công đoạn thu thập, lưu trữ, tìm kiếm, chia sẻ, truyền dẫn, phân tích và hiển thị. Ứng dụng của Big Data gồm có: xác định xu hướng kinh doanh, chất lượng nghiên cứu, ngăn chặn bệnh dịch, liên kết các trích dẫn luật pháp, phòng chống tội phạm, xác định tình trạng giao thông trong thời gian thực... Năm 2001, nhà phân

Hình 1: Mô hình 3Vs của Dữ liệu lớn



Nguồn: Soubra (2012).

tích Doug Laney của hãng META Group (Công ty nghiên cứu Gartner) cho rằng các thách thức và cơ hội nằm trong việc tăng trưởng dữ liệu có thể được mô tả bằng ba chiều "**3Vs**" (Hình 1): số lượng lưu trữ (*volume*), tốc độ xử lý (*velocity*) và chủng loại dữ liệu (*variety*).

Nhiều công ty, tổ chức trong lĩnh vực công nghệ thông tin tiếp tục sử dụng mô hình "**3Vs**" này để định nghĩa Big Data. Năm 2012, Gartner bổ sung ngoài 3 tính chất trên thì Big Data còn phải "*cần đến các dạng xử lý mới để giúp đỡ việc đưa ra quyết định, khám phá sâu vào sự vật/sự việc và tối ưu hóa các quy trình làm việc*". Năm 2014, Gartner có khái niệm mới **"5Vs"** về Big Data. Muốn hiểu và ứng dụng dữ liệu lớn ta phải biết rõ các tính chất của nó. Khái niệm 5Vs của Gartner có thể hiểu như sau:

- *Volume (Dung lượng lưu trữ):* Dữ liệu lớn có dung lượng lưu trữ vượt mức đảm đương của những ứng dụng và công cụ truyền thống. Kích cỡ của nó đang từng ngày tăng lên, có thể nằm trong khoảng vài chục terabyte cho đến nhiều petabyte chỉ cho một tập hợp dữ liệu. *Volume* chỉ độ lớn của dữ liệu ở mức Terabytes (TB), Petabytes (PB), Exabytes (EB), Zettabyte (ZB) hay Yottabyte (YB)…

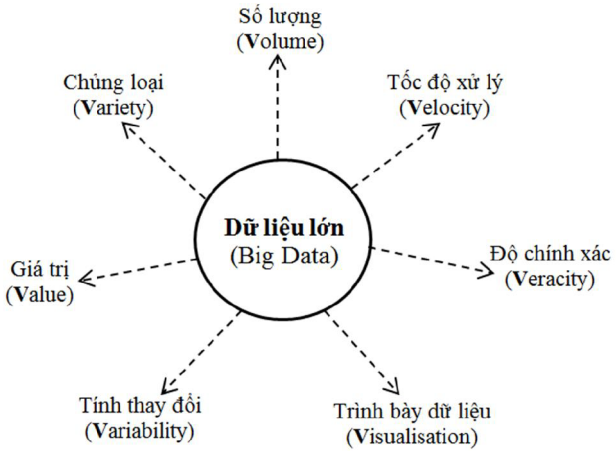
- *Velocity (Tốc độ xử lý):* Dung lượng dữ liệu gia tăng rất nhanh và tốc độ xử lý đang *tiến tới thời gian thực* (real-time). Các ứng dụng phổ biến trên Internet, Tài chính, Ngân hàng, Hàng không, Quân sự, Y tế – Sức khỏe ngày hôm nay phần lớn được xử lý real-time. Công nghệ xử lý dữ liệu lớn ngày một tiên tiến cho phép chúng ta xử lý tức thì trước khi chúng được lưu trữ vào cơ sở dữ liệu. *Velocity* chỉ tính chất chuyển động liên tục của dòng dữ liệu (*Streaming Data*) rất lớn cần xử lý, khác với cách truyền thống ta thu nhận và xử lý dữ liệu theo lô (*Batch*).

- *Variety (Đa dạng chủng loại):* Hình thức lưu trữ và chủng loại dữ liệu ngày một đa dạng hơn. Trước đây ta hay nói đến dữ liệu có cấu trúc thì ngày nay hơn 80% dữ liệu trên thế giới được sinh ra là không cấu trúc (*unstructured data*). Công nghệ Big Data cho phép ta *liên kết và phân tích đa dạng chủng loại dữ liệu* với nhau như *comments* hay *post* của một nhóm người dùng nào đó trên Facebook với thông tin video được chia sẻ từ Youtube và Twitter. *Variety* chỉ sự đa dạng, sự liên kết chằng chịt của dữ liệu với nhiều cấu trúc khác nhau, từ dữ liệu quan hệ đến dữ liệu không cấu trúc như *văn bản*, *blog, hình ảnh, video, audio*...

- *Veracity (Độ chính xác):* Một trong những tính chất phức tạp nhất của Big Data là độ chính xác dữ liệu. Với xu hướng mạng và truyền thông xã hội (*Social Media* và *Social Network*), sự gia tăng mạnh mẽ tính tương tác và chia sẻ của người dùng di động làm cho việc xác định độ tin cậy và chính xác của dữ liệu khó khăn hơn. Do đó yếu tố *Variety* là một tính chất quan trọng của Big Data.

- *Value (Giá trị thông tin):* Đây là tính chất quan

Hình 2: Mô hình 7Vs của Dữ liệu lớn



Nguồn: Tác giả minh họa theo McNulty (2014).

trọng nhất của công nghệ Big Data. Doanh nghiệp phải hoạch định được những giá trị thông tin hữu ích của Big Data cho *vấn đề, bài toán hoặc mô hình hoạt động kinh doanh của mình*. Phải xác định được "*Giá trị của thông tin*" thì mới nên bắt tay vào Big Data. Dữ liệu là nguồn chứa hầu hết mọi thông tin của con người nhưng những thông tin này chỉ có khi chúng được phân tích (xử lý). Xử lý dữ liệu lớn là một vấn đề khó, cho tới nay hầu như chưa có cách làm được tốt nhất việc này.

Hiện nay, ngoài **5V's**, còn có thêm mô hình **7V's** với 2 tính chất được bổ sung là: ***Variability*** - chỉ *sự thay đổi liên tục của dữ liệu*, đặc biệt khi thu thập dữ liệu dựa vào xử lý ngôn ngữ; và ***Visualisation*** – chỉ *cách thức trình bày dữ liệu* làm sao có thể đọc và có thể truy cập được (McNulty, 2014). Về tổng thể đến nay chúng ta có cái nhìn đầy đủ về các tính chất quan trọng của dữ liệu lớn theo mô hình này (Hình 2).

2.2. Các nguồn Dữ liệu lớn

Dữ liệu lớn đang ngày càng đến rất nhiều quanh ta là một hiện thực khách quan. Dữ liệu lớn có ở rất nhiều tổ chức, nhiều hoạt động xã hội, kinh doanh, khoa học và tiềm ẩn nhiều giá trị to lớn. Dữ liệu lớn đến từ rất nhiều nguồn với 3 nguồn chính là: (1) *Các phương tiện truyền thông xã hội,* (2) *Các máy móc thu nhận dữ liệu*, (3) Các g*iao dịch kinh doanh.* Ước tính 100 terabyte dữ liệu được tải lên Facebook mỗi ngày; lượng xử lý của Walmart khoảng 1 triệu giao dịch khách hàng mỗi giờ… Có thể thấy quy mô của dữ liệu là rất lớn. Cùng với Internet di động và sự bùng nổ của thương mại điện tử, hầu hết các ngành đang bị tràn ngập với dữ liệu - chúng có thể sẽ vô cùng quý giá, nếu ta biết cách khai thác, sử dụng nó (McNulty, 2014).

3. Vai trò của dữ liệu lớn trong giáo dục đại học

Nhiều trường đại học bắt đầu đầu tư nhiều hơn trong các dự án dữ liệu lớn và họ được hưởng lợi từ nhiều ứng dụng của nó. Các nhà quản trị giáo dục có không gian dữ liệu hoàn toàn mới để truy cập, cung cấp hiểu biết sâu sắc hơn các hoạt động của trường đại học. Với những công cụ này, viên chức nhà trường có thể tăng cường các khía cạnh khác nhau của cuộc sống đại học. Các nhà giáo dục có thể sử dụng các công cụ dữ liệu lớn để phân tích hiệu quả học tập, kỹ năng của sinh viên nhằm tạo ra một kinh nghiệm học tập được cá nhân hóa đáp ứng nhu cầu cụ thể của sinh viên và phục vụ cho công tác đào tạo, quản lý sinh viên. Các công cụ dữ liệu lớn thậm chí có thể phân tích các nhóm năng động, nhìn vào những điểm mạnh và điểm yếu khác nhau của từng sinh viên để xác định sự sắp xếp tối ưu trong quản lý và đào tạo. Bằng cách tăng cường kinh nghiệm học tập và nâng cao thành tích sinh viên, tăng cường khả năng giảng dạy và nghiên cứu của giảng viên trên diện rộng, các trường đại học sẽ có thể làm giảm tỷ lệ bỏ học, thu hút thêm sinh viên và gia tăng số lượng sinh viên tốt nghiệp, nâng cao vị thế, cải thiện cuộc sống của các thành viên nhà trường. Ngoài những lợi ích đơn thuần về học thuật, các trường có thể mong đợi một tiềm năng lớn hơn từ các cựu sinh viên có khả năng đóng góp cho trường cũ của họ. Một nghiên cứu của Shakil và Faizi (2012) cho thấy vai trò quan trọng của các cựu sinh viên về danh tiếng, tài chính, cả về học thuật lẫn

quản trị tại các trường đại học.

Ngày càng có nhiều nguồn dữ liệu đa dạng cần phải thu thập, lưu trữ và phân tích như: số liệu đăng ký, tuyển sinh, tốt nghiệp của các hệ đào tạo qua các năm; số liệu từ các nhận xét, đánh giá của sinh viên; số liệu từ các cố vấn học tập, các trợ lý Khoa, Viện, đặc biệt là từ hệ thống thông tin quản lý đào tạo; số liệu về các cựu sinh viên và cả các số liệu, báo cáo công bố của nhiều trường và bộ ngành, chính phủ... Theo thời gian, các tập số liệu này càng tăng thêm nên việc lưu trữ, phân tích, xử lý chúng cho các mục tiêu khác nhau trở thành vấn đề lớn. Các trường đại học có đông sinh viên có thể được hưởng lợi từ dữ liệu lớn. Tất cả dữ liệu về sinh viên từ khi nhập học đến khi trở thành các cựu sinh viên là một nguồn dữ liệu lớn hữu ích cho các trường lưu trữ và khai thác. Làm sao để tận dụng các dữ liệu về kết quả học tập, dữ liệu nhân khẩu học, việc làm khi ra trường, hành vi và những hiểu biết xã hội (khuynh hướng, thiên hướng và xu thế) để tối ưu hóa quá trình tham gia của sinh viên, tăng giá trị suốt đời của sinh viên và hướng họ tuyên truyền/giới thiệu về trường? Ta hãy xem xét ví dụ về một số ứng dụng dữ liệu lớn, được hỗ trợ cho tổ chức giáo dục đại học:

• *Thu hút sinh viên:* Sử dụng kết quả và dữ liệu lịch sử của sinh viên hiện tại và các cựu sinh viên để tìm ra hồ sơ các ứng viên có nhiều khả năng đăng ký, sau đó bổ sung thêm với các dữ liệu truyền thông xã hội để ghi nhận những ưu tiên trong lựa chọn trường. Dùng các phân tích để tìm hiểu các mạng xã hội của sinh viên hiện tại và tương lai để xác định bạn bè, người thân của họ có thể là những sinh viên tiềm năng mới.

• *Lựa chọn chuyên ngành, môn học:* Tạo các hồ sơ chi tiết dựa trên kết quả học tập ở bậc học phổ thông, các lĩnh vực quan tâm được nắm bắt trong các khảo sát trên phương tiện truyền thông xã hội, trên kết quả kiểm tra năng lực đầu vào. So sánh các hồ sơ về các khóa học và các chuyên ngành để tìm ra sự sắp xếp, lựa chọn phù hợp. Việc tích hợp với các dữ liệu bên ngoài liên quan đến các nhu cầu về kỹ năng và thu nhập của nguồn nhân lực trong tương lai sẽ giúp sinh viên đưa ra các quyết định lựa chọn.

• *Đánh giá hiệu quả học tập của sinh viên:* Giám sát kết quả thi của sinh viên hiện tại và so sánh với các kết quả thi trước đó. Tích hợp dữ liệu truyền thông xã hội và các ghi chép, đánh giá của giảng viên để tạo ra hồ sơ chi tiết về hành vi và khuynh hướng của sinh viên. Khai triển các đề nghị cụ thể của sinh viên, chẳng hạn như việc hướng dẫn cá nhân hay các nhóm nhỏ sinh viên, các tài liệu bổ sung trong các môn học, hoặc thậm chí thay đổi trong các lớp hoặc chuyên ngành.

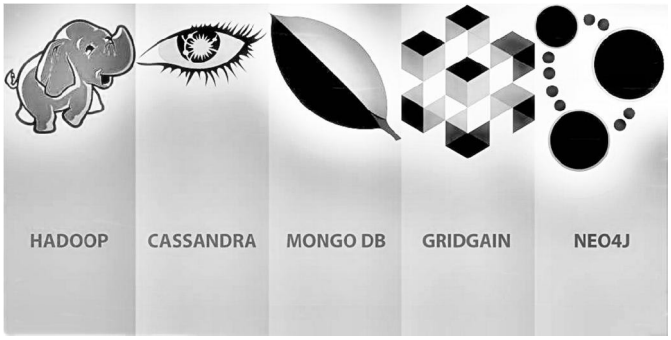
• *Các nhóm sinh viên:* Phân tích phân nhóm (*cohort analysis*) cho các nhóm sinh viên (*student workgroups*) có thể làm việc nhóm trong và ngoài lớp học để cải thiện hiệu suất từng cá nhân. Các phân tích cho phép giảng viên xem xét sự phân nhóm, các yếu tố và lý do cho việc phân nhóm và cho phép giảng viên hủy bỏ. Các phân tích có thể cải tổ phân nhóm dựa trên các yếu tố đã thiết kế hoặc các yếu tố ngẫu nhiên, giảng viên lưu lại các quan sát sau khi cải tổ để cập nhật tập dữ liệu.

• *Giữ chân sinh viên:* Kết hợp các phân tích và thang điểm trước đó cả về hiệu quả hoạt động của sinh viên và các nhóm sinh viên, kết hợp với dữ liệu về nhân khẩu học, dữ liệu tài chính và xã hội để đánh giá khả năng hao hụt, cho phép cơ sở giáo dục đưa ra quyết định có nên giữ lại sinh viên này hay không. Cung cấp và đo lường hiệu quả của các đề nghị cụ thể, dựa trên sự thành công của biện pháp can thiệp trước đó. Trao quyền cho giảng viên thực hiện các đề nghị của họ, các đề nghị này có thể được theo dõi kết quả và áp dụng các đề xuất giữ lại trong tương lai.

• *Hiệu quả của giảng viên:* Đo lường và tinh chỉnh hiệu suất giảng viên. Trong khi một số cơ sở giáo dục bị giới hạn hiệu suất có thể hưởng lợi từ cái nhìn sâu sắc vào hiệu quả một giảng viên khi so sánh tương tự giữa các giảng viên. Hiệu suất có thể được xác định theo chủ đề, số lượng sinh viên, dữ liệu nhân khẩu học, các phân loại hành vi, nguyện vọng của sinh viên và một số yếu tố khác để đảm bảo rằng giảng viên phù hợp với đúng lớp học và sinh viên để đảm bảo tốt nhất cho giảng viên cũng như sinh viên. Từ phản hồi của sinh viên, giảng viên có điều kiện rút kinh nghiệm, cải tiến và nâng cao chất lượng giảng dạy (Allouche, 2014).

• *Giá trị suốt đời của sinh viên:* Lập kế hoạch liên quan đến tiềm năng cho, tặng của sinh viên hiện tại và cựu sinh viên. Hiểu biết khả năng và tiềm năng về thu nhập, tài sản hiện tại hoặc tương lai có thể là những yếu tố chính trong hồ sơ để tối ưu hóa các

Hình 3: Năm công cụ dữ liệu lớn mã nguồn mở hàng đầu



Nguồn: Panigrahi (2014).

nguồn lực mà cựu sinh viên có thể tài trợ. Tận dụng lợi thế của các hiểu biết này để xác định sớm những người ủng hộ trong tương lai và tiên đoán hiệu suất cao nhất người ủng hộ. Trên thực tế, việc duy trì, cập nhật thông tin và tận dụng mạng lưới cựu sinh viên là rất hữu ích, là mong muốn của nhiều đơn vị cần sự hỗ trợ của các cựu sinh viên thành đạt, song lại là một công việc khó khăn.

• *Sự ủng hộ tích cực của sinh viên:* Tận dụng các phân tích sinh động bằng đồ thị hay biểu đồ nhằm đánh giá, theo dõi khả năng để đề xuất và chọn những lĩnh vực cụ thể về những trải nghiệm tại trường học là tốt hay tồi hơn đối với một sinh viên riêng biệt. Sử dụng dữ liệu này để tạo ra thang điểm đánh giá sự ủng hộ của sinh viên có thể được tận dụng trong việc thu hút sinh viên, duy trì sinh viên, hiệu quả học tập của sinh viên và các ứng dụng về giá trị suốt đời của sinh viên.

• *Thúc đẩy giáo dục với dữ liệu nguồn mở:* Một điểm hay của việc phân tích trong không gian giáo dục là nhóm các dữ liệu mà các tổ chức giáo dục hay Nhà nước nắm giữ và công bố. Các dữ liệu từ các báo cáo này có thể được tích hợp với dữ liệu của đơn vị để tạo ra các tiêu chuẩn, dựa vào đó ta có thể so sánh hiệu quả và xác định các lĩnh vực tiềm năng của hoạt động quản lý.

4. Một số công cụ dữ liệu lớn

Các tổ chức đang phải đối mặt với thách thức để quản lý và sử dụng dữ liệu lớn một cách hiệu quả nhất. Các công cụ mã nguồn mở giúp các tổ chức xử lý số lượng lớn các dữ liệu mà họ có được dễ dàng hơn. Ở đây liệt kê 5 công cụ dữ liệu lớn mã nguồn mở hàng đầu (Hình 3).

Hadoop: Bất kỳ cuộc thảo luận nào về dữ liệu lớn chắc chắn phải đề cập đến Hadoop. Apache Hadoop là một phần mềm xử lý dữ liệu rất hiệu quả, hỗ trợ nhiều hệ điều hành.

Cassandra: Được quỹ Apache quản lý, nhưng nó lại được Facebook phát triển. Cơ sở dữ liệu NoSQL (*Not Relational SQL* hay *Not Only SQL*) này tiện dụng cho các tổ chức làm việc với các tập dữ liệu rất lớn.

MongoDB: Cơ sở dữ liệu NoSQL này được tạo ra để quản lý dữ liệu khổng lồ. Tính năng gồm: hỗ trợ đánh chỉ số đầy đủ, lưu trữ tài liệu theo định hướng, nhân rộng, tính sẵn sàng cao.

GridGain: như là một thay thế cho MapReduce của Google. Nó tương thích với Hadoop Distributed File System.

Neo4j: được cho là cơ sở dữ liệu đồ thị tốt nhất. Neo4j sử dụng trong các ứng dụng quan trọng của các doanh nghiệp và chính phủ trên khắp thế giới.

5. Kết luận

Có vô số các cơ hội cho các cơ sở giáo dục sử dụng dữ liệu lớn để cải thiện sự hài lòng và hiệu quả hoạt động của sinh viên, giảng viên, của công tác quản lý. Mặc dù có sự cạnh tranh giữa các trường trong việc tuyển sinh, tất cả có thể được hưởng lợi từ việc chia sẻ dữ liệu, phân tích và áp dụng tốt nhất trong các lĩnh vực như đánh giá hiệu quả học tập của sinh viên, duy trì số lượng sinh viên theo học, giữ chân các giảng viên giỏi và nhiều hơn nữa. Trong một thế giới mà giáo dục nắm giữ tiềm năng lớn nhất để hướng đến chất lượng cuộc sống, rất nhiều cơ hội cho các tổ chức giáo dục hợp tác nhằm mang lại nhiều lợi ích cho sinh viên, đội ngũ giảng viên và cả xã hội nói chung. Việc áp dụng công nghệ dữ liệu lớn trong hệ thống các trường đại học sẽ mất thời gian như: nhiều trường sẽ đối mặt với nguồn lực hạn chế, mối bận tâm về vấn đề bảo mật và tính riêng tư

sẽ cần phải được giải quyết triệt để. Tuy nhiên, khi tìm thấy sự cân bằng, dữ liệu lớn có thể chứng minh là một nguồn tài nguyên vô giá cho các trường nghiên cứu, khai thác hiệu quả. Để giải quyết các vấn đề trên, tác giả cho rằng các trường đại học ở nước ta cần hoạch định và tiến hành một số việc sau:

1) *Các nhà quản lý giáo dục cần nhận thức được ý nghĩa vai trò của tất cả các nguồn dữ liệu nội tại và xung quanh môi trường giáo dục*. Đó là nguồn dữ liệu lớn phải được thu thập, lưu trữ, xử lý và phân tích để phục vụ tốt hơn các nhiệm vụ trọng tâm của tổ chức. "Tin học hóa" trong các nhà trường không chỉ là mua sắm, trang bị; nó chỉ là công cụ và chỉ phát huy hiệu quả khi biết tận dụng khai thác nguồn lực hiện có và tập trung vào mục tiêu mà tổ chức đang cần. Các báo cáo, nghiên cứu, các hội thảo khoa học về dữ liệu lớn sẽ tăng cường thêm nhận thức và hành động cụ thể trong lĩnh vực này.

2) *Từ chỗ ý thức được vấn đề, cần chuẩn bị nguồn nhân lực cho lĩnh vực dữ liệu lớn*. Nhiều cơ quan, tổ chức đang cố gắng để hiểu về Big Data và chuẩn bị lực lượng lao động cho thời đại Big Data. Ngay từ 2011, McKinsey Global Institute đã dự đoán về sự thiếu hụt các nhà khoa học và quản trị dữ liệu lớn (Manyika và cộng sự, 2011). Do đó, các tổ chức sẽ ngày càng khuyến khích nhân viên có tay nghề cao tìm hiểu nhiều hơn nữa về Big Data. Bên cạnh đó, sinh viên mới tốt nghiệp hoặc sinh viên hiểu rõ về những xu hướng dữ liệu lớn sẽ có nhiều cơ hội cạnh tranh trên thị trường việc làm. Nhìn vào danh mục tìm người làm trên tạp chí Wall Street Journal, ta có thể thấy các công ty thương mại đều đang tìm thuê các nhà khoa học dữ liệu (*Data Scientists*) và phân tích viên dữ liệu (*Data analyst*) đến làm việc và sự cạnh tranh là dữ dội.

3) *Xác định và dành cho lĩnh vực này nguồn lực tài chính, vật chất để có thể triển khai trên thực tế*, nếu không mọi thứ sẽ vẫn chỉ là ý tưởng.

4) *Tổ chức, sắp xếp lại nhân sự, quy trình, thủ tục một cách chặt chẽ, nhất quán* trong chuỗi các khâu thu thập, lưu trữ, xử lý và phân tích các nguồn dữ liệu đa dạng để có được các thông tin hữu dụng cho công tác quản trị đại học.

5) *Tiếp cận, làm quen, tập huấn và đào tạo nguồn nhân sự* sử dụng khai thác các công cụ dữ liệu lớn, trước hết là tận dụng các công cụ nguồn mở để sẵn sàng triển khai các dự án dữ liệu lớn.

6) *Liên kết, chia sẻ các nguồn lực thông tin* giữa các cơ quan, tổ chức giáo dục và các tổ chức khác trong khai thác ứng dụng dữ liệu lớn để đạt hiệu quả cao hơn.

Dữ liệu lớn và công nghệ có lẽ sẽ không phải là thuốc chữa bách bệnh cho tất cả các tồn tại của giáo dục đại học. Điều quan trọng là xây dựng công nghệ với nhận thức về vai trò và tác động của nó nhằm phục vụ hiệu quả mục đích cuối cùng của tổ chức. Cũng như với tất cả các ngành khác, cần có sự hiểu biết, áp dụng các dữ liệu và các quá trình phân tích để giáo dục mang lại kết quả tích cực, đem lại lợi ích cho cả giảng viên, sinh viên và các nhà quản lý. Với khuôn khổ bài báo, tác giả giới thiệu và phân tích một số vấn đề liên quan đến vai trò và ảnh hưởng của công nghệ dữ liệu lớn trong giáo dục đại học nói chung. Các tổ chức giáo dục đại học ở Việt Nam quan tâm đến dữ liệu lớn cần đi sâu phân tích, tìm hiểu để có được giải pháp ứng dụng cụ thể trong hoàn cảnh của mình.❐

Tài liệu tham khảo

Allouche, G. (2014), *Understanding how Big Data can Help Improve Teaching, With 2 Examples,* truy cập ngày 12 tháng 6 năm 2014, từ <http://www.emergingedtech.com/2014/06/how-big-data-can-help-improve-teaching/>.

Gartner (2013), *Gartner Special Report Examines Trends in Big Data*, truy cập ngày 12 tháng 3 năm 2013, từ <http://www.gartner.com/newsroom/id/2366515>.

Laney, D. (2001), *3-D Data Management: Controlling Data Volume, Velocity and Variety,* META Group, truy cập tháng 1 năm 2012, từ <http://blogs.gartner.com/doug-laney/deja-vvvue-others-claiming-gartners-volume-velocity-variety-construct-for-big-data/>.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. và Byers, Angela Hung (2011), "*Big data: The next frontier for innovation, competition, and productivity"*, truy cập tháng 5 năm 2011, từ <http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation>.

Marr, B. (2013), *The Awesome Ways Big Data Is Used Today To Change Our World*, truy cập tháng 11 năm 2013, từ <http://kpilibrary.com/topics/10-awesome-ways-big-data-is-used-today-to-change-our-world>.

Marr, B. (2014), *Big Data: The 5 Vs Everyone Must Know*, truy cập tháng 3/2014, từ <https://www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know>.

Mashey, R.J. (1998), *Big Data… and the Next Wave of Infrastress, A Quantitiative Approach, Second Edition*, Morgan Kaufmann, San Francisco, truy cập tháng 9 năm 2014, từ < http://www.usenix.org/event/usenix99/invited_talks/mashey.pdf>

McNulty, E. (2014), *Understanding Big Data: The Seven V's,* truy cập tháng 5 năm 2014, từ <http://dataconomy.com/seven-vs-big-data/>.

Nagel, D. (2013), *IT Investments in Big Data Increasing Across the Board*, truy cập ngày 12 tháng 3 năm 2013, từ <http://campustechnology.com/Articles/2013/03/12/IT-Investments-in-Big-Data-Increasing-Across-the-Board.aspx?Page=1>.

Panigrahi, *S. (2014), Top 5 open source tools for big data, truy cập tháng 5 năm 2014, từ <https://www.linkedin.com/pulse/20140530053319-19479208-top-5-open-source-tools-for-big-data>.*

Press, G. (2013), *A Very Short History of Big Data*, truy cập tháng 12 năm 2013, từ <http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/>.

Schmarzo, B. (2014), *What Universities Can Learn from Big Data – Higher Education Analytics,* EMC Corporation, 2014.

Shakil, A. và Faizi, W. (2012), "*The Importance of Alumni Association at University Levelin Karachi, Pakistan", Tạp chí Education,* số 2(1), tr. 20-25*,* truy cập tháng 12 năm 2014, từ cơ sở dữ liệu Academia.

Soubra, D. (2012), *The 3Vs that define Big Data*, truy cập ngày 5 tháng 7 năm 2012, từ

<http://www.datasciencecentral.com/forum/topics/the-3vs-that-define-big-data>.

Vu, J. (2014), *Xu hướng Big Data,* Blogs của Giáo sư John Vũ, Đại học Carnegie Mellon, truy cập tháng 2 năm 2014, từ <http://science-technology.vn/?p=4101>.

White, T. (2012), *Hadoop: The Definitive Guide, 3rd Edition,* Nhà xuất bản O'Reilly Media, California, 2012.

Thông tin tác giả:

***Vũ Hưng Hải,** *thạc sỹ*

- Tổ chức tác giả công tác: Khoa Tin học Kinh tế, Đại học Kinh tế quốc dân

- Lĩnh vực nghiên cứu chính: Hệ thống thông tin, An toàn và bảo mật thông tin.

- Địa chỉ liên hệ: Địa chỉ Email: haivh@neu.edu.vn