Chapter 3
The Big Data Value Chain: Definitions, Concepts, and Theoretical Approaches

Edward Curry

3.1 Introduction

The emergence of a new wave of data from sources, such as the Internet of Things, Sensor Networks, Open Data on the Web, data from mobile applications, social network data, together with the natural growth of datasets inside organisations (Manyika et al. 2011), creates a demand for new data management strategies which can cope with these new scales of data environments. Big data is an emerging field where innovative technology offers new ways to reuse and extract value from information. The ability to effectively manage information and extract knowledge is now seen as a key competitive advantage, and many organisations are building their core business on their ability to collect and analyse information to extract business knowledge and insight. Big data technology adoption within industrial sectors is not a luxury but an imperative need for most organisations to gain competitive advantage.

This chapter examines definitions and concepts related to big data. The chapter starts by exploring the different definitions of "Big Data" which have emerged over the last number of years to label data with different attributes. The Big Data Value Chain is introduced to describe the information flow within a big data system as a series of steps needed to generate value and useful insights from data. The chapter explores the concept of Ecosystems, its origins from the business community, and how it can be extended to the big data context. Key stakeholders of a big data ecosystem are identified together with the challenges that need to be overcome to enable a big data ecosystem in Europe.

E. Curry (✉)
Insight Centre for Data Analytics, National University of Ireland Galway, Lower Dangan, Galway, Ireland
e-mail: edward.curry@insight-centre.org

J.M. Cavanillas et al. (eds.), *New Horizons for a Data-Driven Economy*,
DOI 10.1007/978-3-319-21569-3_3

3.2 What Is Big Data?

Over the last years, the term "Big Data" was used by different major players to label data with different attributes. Several definitions of big data have been proposed over the last decade; see Table 3.1. The first definition, by Doug Laney of META Group (then acquired by Gartner), defined big data using a three-dimensional perspective: "Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision-making, insight discovery and process optimization" (Laney 2001). Loukides (2010) defines big data as "when the size of the data itself becomes part of the problem and traditional techniques for working with data run out of steam". Jacobs (2009) describes big data as "data whose size forces us to look beyond the tried-and-true methods that are prevalent at that time".

Big data brings together a set of data management challenges for working with data under new scales of size and complexity. Many of these challenges are not new. What is new however are the challenges raised by the specific characteristics of big data related to the 3 Vs:

- **Volume (amount of data)**: dealing with large scales of data within data processing (e.g. Global Supply Chains, Global Financial Analysis, Large Hadron Collider).
- **Velocity (speed of data):** dealing with streams of high frequency of incoming real-time data (e.g. Sensors, Pervasive Environments, Electronic Trading, Internet of Things).
- **Variety (range of data types/sources):** dealing with data using differing syntactic formats (e.g. Spreadsheets, XML, DBMS), schemas, and meanings (e.g. Enterprise Data Integration).

The Vs of big data challenge the fundamentals of existing technical approaches and require new forms of data processing to enable enhanced decision-making, insight discovery, and process optimisation. As the big data field matured, other Vs have been added such as Veracity (documenting quality and uncertainty), Value, etc. The value of big data can be described in the context of the dynamics of knowledge-based organisations (Choo 1996), where the processes of decision-making and organisational action are dependent on the process of sense-making and knowledge creation.

3.3 The Big Data Value Chain

Within the field of Business Management, Value Chains have been used as a decision support tool to model the chain of activities that an organisation performs in order to deliver a valuable product or service to the market (Porter 1985). The value chain categorises the generic value-adding activities of an organisation

Table 3.1 Definitions of big data

Big data definition	Source
"Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization"	Laney (2001), Manyika et al. (2011)
"When the size of the data itself becomes part of the problem and traditional techniques for working with data run out of steam"	Loukides (2010)
Big Data is "data whose size forces us to look beyond the tried-and-true methods that are prevalent at that time"	Jacobs (2009)
"Big Data technologies [are] a new generation of technologies and architectures designed to extract value economically from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis"	IDC (2011)
"The term for a collection of datasets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications"	Wikipedia (2014)
"A collection of large and complex data sets which can be processed only with difficulty by using on-hand database management tools"	Mike 2.0 (2014)
"Big Data is a term encompassing the use of techniques to capture, process, analyse and visualize potentially large datasets in a reasonable timeframe not accessible to standard IT technologies." By extension, the platform, tools and software used for this purpose are collectively called "Big Data technologies"	NESSI (2012)
"Big data can mean big volume, big velocity, or big variety"	Stonebraker (2012)

allowing them to be understood and optimised. A value chain is made up of a series of subsystems each with inputs, transformation processes, and outputs. Rayport and Sviokla (1995) were one of the first to apply the value chain metaphor to information systems within their work on Virtual Value Chains. As an analytical tool, the value chain can be applied to information flows to understand the value creation of data technology. In a Data Value Chain, information flow is described as a series of steps needed to generate value and useful insights from data. The European Commission sees the data value chain as the "centre of the future knowledge economy, bringing the opportunities of the digital developments to the more traditional sectors (e.g. transport, financial services, health, manufacturing, retail)" (DG Connect 2013).

The Big Data Value Chain (Curry et al. 2014), as illustrated in Fig. 3.1, can be used to model the high-level activities that comprise an information system. The Big Data Value Chain identifies the following key high-level activities:

Data Acquisition is the process of gathering, filtering, and cleaning data before it is put in a data warehouse or any other storage solution on which data analysis can be carried out. Data acquisition is one of the major big data challenges in terms of infrastructure requirements. The infrastructure required to support the acquisition of big data must deliver low, predictable latency in both capturing data and in executing queries; be able to handle very high transaction volumes, often in a

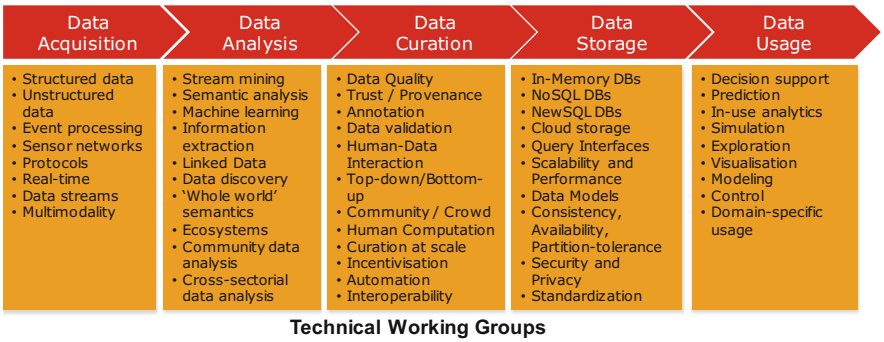

Fig. 3.1 The Big Data Value Chain as described within (Curry et al. 2014)

distributed environment; and support flexible and dynamic data structures. Data acquisition is further detailed in this chapter.

Data Analysis is concerned with making the raw data acquired amenable to use in decision-making as well as domain-specific usage. Data analysis involves exploring, transforming, and modelling data with the goal of highlighting relevant data, synthesising and extracting useful hidden information with high potential from a business point of view. Related areas include data mining, business intelligence, and machine learning. Chapter 4 covers data analysis.

Data Curation is the active management of data over its life cycle to ensure it meets the necessary data quality requirements for its effective usage (Pennock 2007). Data curation processes can be categorised into different activities such as content creation, selection, classification, transformation, validation, and preservation. Data curation is performed by expert curators that are responsible for improving the accessibility and quality of data. Data curators (also known as scientific curators, or data annotators) hold the responsibility of ensuring that data are trustworthy, discoverable, accessible, reusable, and fit their purpose. A key trend for the curation of big data utilises community and crowd sourcing approaches (Curry et al. 2010). Further analysis of data curation techniques for big data is provided in Chap. 5.

Data Storage is the persistence and management of data in a scalable way that satisfies the needs of applications that require fast access to the data. Relational Database Management Systems (RDBMS) have been the main, and almost unique, solution to the storage paradigm for nearly 40 years. However, the ACID (Atomicity, Consistency, Isolation, and Durability) properties that guarantee database transactions lack flexibility with regard to schema changes and the performance and fault tolerance when data volumes and complexity grow, making them unsuitable for big data scenarios. NoSQL technologies have been designed with the scalability goal in mind and present a wide range of solutions based on alternative data models. A more detailed discussion of data storage is provided in Chap. 6.

Data Usage covers the data-driven business activities that need access to data, its analysis, and the tools needed to integrate the data analysis within the business activity. Data usage in business decision-making can enhance competitiveness through reduction of costs, increased added value, or any other parameter that can be measured against existing performance criteria. Chapter 7 contains a detailed examination of data usage.

3.4 Ecosystems

The term *ecosystem* was coined by Tansley in 1935 to identify a basic ecological unit comprising of both the environment and the organisms that use it. Within the context of business, James F. Moore (1993, 1996, 2006) exploited the biological metaphor and used the term to describe the business environment. Moore defined a business ecosystem as an "economic community supported by a foundation of interacting organizations and individuals" (Moore 1996). A strategy involving a company attempting to succeed alone has proven to be limited in terms of its capacity to create valuable products or services. It is crucial that businesses collaborate among themselves to survive within a business ecosystem (Moore 1993; Gossain and Kandiah 1998). Ecosystems allow companies to create new value that no company could achieve by itself (Adner 2006). Within a healthy business ecosystem, companies can work together in a complex business web where they can easily exchange and share vital resources (Kim et al. 2010).

The study of Business Ecosystems is an active area of research where researchers are investigating many facets of the business ecosystem metaphor to explore aspects such as community, cooperation, interdependency, co-evolution, eco-systemic functions, and boundaries of business environments. Koening (2012) provides a simple typology of Business Ecosystems based on the degree of key resource control and type of member interdependence. Types of business ecosystems include supply systems (i.e. Nike), platforms (Apple iTunes), communities of destiny (i.e. Sematech in the semiconductor industry), and expanding communities.

3.4.1 Big Data Ecosystems

In natural ecosystems, smart organisms control their energy. In business ecosystems, a smart company manages information and its flows (Kim et al. 2010). In terms of data, the ecosystem metaphor is useful to describe the data environment supported by a community of interacting organisations and individuals. Big Data Ecosystems can form in different ways around an organisation, community technology platforms, or within or across sectors. Big Data Ecosystems exist within many industrial sectors where vast amount of data move between actors within complex information supply chains. Sectors with established or emerging data

ecosystems include Healthcare, Finance (O'Riáin et al. 2012), Logistics, Media, Manufacturing, and Pharmaceuticals (Curry et al. 2010). In addition to the data itself, Big Data Ecosystems can also be supported by data management platforms, data infrastructure (e.g. Various Apache open source projects), and data services.

3.4.2 European Big Data Ecosystem

While no coherent data ecosystem exists at the European-level (DG Connect 2013), the benefits of sharing and linking data across domains and industry sectors are becoming obvious. Initiatives such as smart cities are showing how different sectors (i.e. energy and transport) can collaborate to maximise the potential for optimisation and value return. The cross-fertilisation of stakeholder and datasets from different sectors is a key element for advancing the big data economy in Europe.

A European big data business ecosystem is an important factor for commercialisation and commoditisation of big data services, products, and platforms. A successful big data ecosystem would see all "stakeholders interact seamlessly within a Digital Single Market, leading to business opportunities, easier access to knowledge and capital" (European Commission 2014).

A well-functioning working data ecosystem must bring together the key stakeholders with a clear benefit for all. The key actors in a big data ecosystem, as illustrated in Fig. 3.2, are:

- **Data Suppliers:** Person or organisation [Large and small and medium-sized enterprises (SME)] that create, collect, aggregate, and transform data from both public and private sources
- **Technology Providers:** Typically organisations (Large and SME) as providers of tools, platforms, services, and know-how for data management
- **Data End Users:** Person or organisation from different industrial sectors (private and public) that leverage big data technology and services to their advantage.
- **Data Marketplace:** Person or organisation that host data from publishers and offer it to consumers/end users.
- **Start-ups and Entrepreneurs:** Develop innovative data-driven technology, products, and services.
- **Researchers and Academics:** Investigate new algorithms, technologies, methodologies, business models, and societal aspects needed to advance big data.
- **Regulators** for data privacy and legal issues.
- **Standardisation Bodies:** Define technology standards (both official and de facto) to promote the global adoption of big data technology.
- **Investors, Venture Capitalists, and Incubators:** Person or organisation that provides resources and services to develop the commercial potential of the ecosystem.

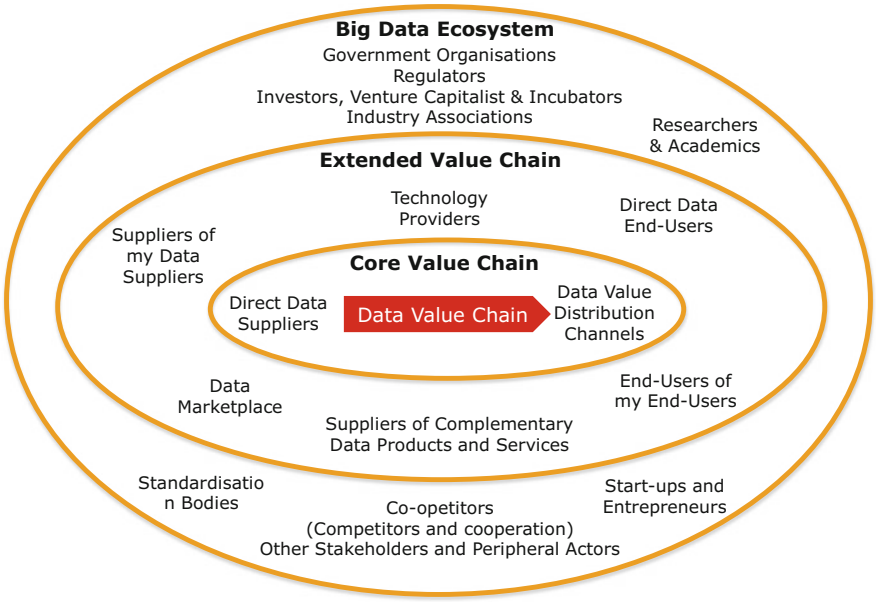

Fig. 3.2 The Micro, Meso, and Macro Levels of a Big Data Ecosystem [adapted from Moore (1996)]

3.4.3 Toward a Big Data Ecosystem

Enabling a European wide data ecosystem will require a number of technical challenges to be overcome associated with the cost and complexity of publishing and utilising data. Current ecosystems face a number of problems such as data discovery, curation, linking, synchronisation, distribution, business modelling, and sales and marketing. A number of key societal and environmental challenges need to be overcome to establish effective big data ecosystems; these include but are not limited to:

- Understanding the value and contribution of big data technology
- Determining the value of data
- Identification of business models that will support a data-driven ecosystem
- Enabling entrepreneurs and venture capitalists to easily access the ecosystem
- Preservation of privacy and security for all actors in the ecosystem
- Reducing fragmentation of languages, intellectual property rights, laws, and policy practices between EU countries

3.5 Summary

Big data is the emerging field where innovative technology offers new ways to extract value from the tsunami of available information. As with any emerging area, terms and concepts can be open to different interpretations. The Big Data domain is no different. The different definitions of "Big Data" which have emerged show the diversity and use of the term to label data with different attributes. Two tools from the business community, Value Chains and Business Ecosystems, can be used to model big data systems and the big data business environments. Big Data Value Chains can describe the information flow within a big data system as a series of steps needed to generate value and useful insights from data. Big Data Ecosystems can be used to understand the business context and relationships between key stakeholders. A European big data business ecosystem is an important factor for commercialisation and commoditisation of big data services, products, and platforms.

References

Adner, R. (2006). Match your innovation strategy to your innovation ecosystem. *Harvard Business Review, 84*, 98–107.

Choo, C. W. (1996). The knowing organization: How organizations use information to construct meaning, create knowledge and make decisions. *International Journal of Information Management, 16*, 329–340. doi:10.1016/0268-4012(96)00020-5.

Curry, E., Ngonga, A., Domingue, J., Freitas, A., Strohbach, M., Becker, T., et al. (2014). D2.2.2. Final version of the technical white paper. Public deliverable of the EU-Project BIG (318062; ICT-2011.4.4).

Curry, E., Freitas, A., & O'Riáin, S. (2010). The role of community-driven data curation for enterprises. In D. Wood (Ed.), *Linking enterprise data* (pp. 25–47). Boston, MA: Springer US.

DG Connect. (2013). *A European strategy on the data value chain*.

European Commission. (2014). *Towards a thriving data-driven economy, Communication from the commission to the European Parliament, the council, the European economic and social Committee and the committee of the regions*, Brussels.

Gossain, S., & Kandiah, G. (1998). Reinventing value: The new business ecosystem. *Strategy and Leadership, 26*, 28–33.

IDC. (2011). *IDC's worldwide big data taxonomy*.

Jacobs, A. (2009). The pathologies of big data. *Communications of the ACM, 52*, 36–44. doi:10.1145/1536616.1536632.

Kim, H., Lee, J.-N., & Han, J. (2010). The role of IT in business ecosystems. *Communications of the ACM, 53*, 151. doi:10.1145/1735223.1735260.

Koenig, G. (2012). Business ecosystems revisited. *Management, 15*, 208–224.

Laney, D. (2001). *3D data management: Controlling data volume, velocity, and variety*. Technical report, META Group.

Loukides, M. (2010). What is data science? *O'Reily Radar*.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute, p. 156.

Mike 2.0. (2014). Big data definition – Mike 2.0.

Moore, J. F. (1993). Predators and prey: A new ecology of competition. *Harvard Business Review, 71*, 75–86.

Moore, J. F. (1996). *The death of competition: Leadership and strategy in the age of business ecosystems*. New York: HarperCollins.

Moore, J. F. (2006). Business ecosystems and the view from the firm. *Antitrust Bulletin, 51*, 31–75.

NESSI. (2012). Big data: A new world of opportunities. NESSI White Paper.

O'Riáin, S., Curry, E., & Harth, A. (2012). XBRL and open data for global financial ecosystems: A linked data approach. *International Journal of Accounting Information Systems, 13*, 141–162. doi:10.1016/j.accinf.2012.02.002.

Pennock, M. (2007). Digital curation: A life-cycle approach to managing and preserving usable digital information. *Library and Archives Journal, 1*, 1–3.

Porter, M. E. (1985). *Competitive advantage: Creating and sustaining superior performance*. New York: Free Press. doi:10.1182/blood-2005-11-4354.

Rayport, J. F., & Sviokla, J. J. (1995). Exploiting the virtual value chain. *Harvard Business Review, 73*, 75–85. doi:10.1016/S0267-3649(00)88914-1.

Stonebraker, M. (2012). What does 'big data' mean. *Communications of the ACM*, BLOG@ ACM.

Tansley, A. G. (1935). The use and abuse of vegetational concepts and terms. *Ecology, 16*, 284–307.

Wikipedia. (2014) Big data. Wikipedia article. http://en.wikipedia.org/wiki/Big_data