Big Data Revolution – What Is It?

Tuomo Nikulainen is a Research Economist at Etlatieto Oy, a subsidiary of ETLA, and currently a Visiting Scholar at University of California, Berkeley (tuomo.nikulainen@etla.fi).

The ongoing ICT revolution has created a vast digital ecosystem in which an increasingly large quantity of information is produced, collected and stored. This development is the result of the digitalization of data and the decreasing cost of data processing and storage. The ability of firms to extract new insights and innovations from massive amounts of data offers significant opportunities to increase productivity and reduce in manufacturing and service costs. This is the next stage of the ICT revolution, and the radical nature of this change is becoming obvious. We are entering the era of data – Big Data.

It is argued that Big Data will provide opportunities with significant financial value across the world and across various sectors. For example, McKinsey[1] estimates the potential for the following:

– cost reduction in US healthcare ($300 billion/year)
– cost reduction in European public sector administration (€250 billion/year)
– increases in net margins in US retail (60% increase)
– decreases in product development and assembly costs for manufacturing (50% decrease)

These estimates suggest that there are significant data-based opportunities in both the private and public sectors. It is important to understand the reasoning behind these projections and who will benefit from them.

The aim of this brief is to discuss the use of Big Data and data analytics in redefining business strategies and decision-making processes. We focus on the characteristics of Big Data, its uses, its potential users, the business options for implementing its data-driven activities, and the role of Big Data in Finland.

What is Big Data?

Big Data is typically characterized by the three V's.

– *Volume* refers to the amount of digitalized data, which has grown exponentially in the last years. More data have been created during the last decade than have been produced over all of the preceding human history.

– *Velocity* comes from the increasing speed of data accumulation. Every day, firms collect enormous quantities of data, and the total amount of data stored by large firms may exceed a petabyte (= 1 million gigabytes).

– *Variability* refers to the different types of data being generated. Traditional business data include financial and marketing data, which come in structured formats with easily identifiable relations between different data points. However, an increasing amount of the data collected comes in unstructured formats (such as web logs, text, images, and video) in which relations between data points are complex. In addition, these data come from a variety of different sources (e.g., internal data generation, internet, mobile), and the number of data-gen-

erating platforms may increase even more significantly in the near future with the potential introduction of smart glasses, smart watches, and advanced sensor networks that will form the core of the next-generation Internet. In developed countries, most people are already walking data generators because smartphones constantly collect information about geographical locations and data usage.

The combination of the three V's stretches the limits of the existing data management and analytics solutions. This limitation relates to the data format that existing systems can handle. The data that can be analyzed in existing data management solutions must be in a structured format in which relations between different data points are well defined. When the data are unstructured and complex (i.e. without structure and without clear relations), new data management and analytics solutions are required. Advanced data solutions are required to create relations for the complex unstructured data and thus make the data usable for analytics.

How is Big Data managed and analyzed?

Data management technologies and techniques related to Big Data come in many shapes and forms. The same holds true for analytics solutions. Different combinations of these techniques and technologies are used in advanced data management and analytics (Table 1).

Example: Facebook and Big Data

Volume
Facebook's main database is approximately 150 petabytes (1 petabyte = 1 thousand terabytes = 1 million gigabytes). This does not include pictures, which are stored in a separate database.

Velocity
Facebook collects 50 terabytes (1 terabyte = 1000 gigabytes) of data every day from its 600 million daily users; its monthly users number over 1 billion.

Variability
Facebook's data cover a variety of formats and domains, such as text, log information, pictures, user-to-user interactions, messages, ad usage, purchases, games, and geo-location.

Broadly speaking, Big Data refers to data management and analytics solutions that are designed for large and complex datasets that are difficult to process with traditional database management tools and data processing applications. Thus, Big Data signifies both the large volume of data and the technologies that enable advanced data management and analytics.

We are also witnessing the rapid decline of computing and data storage costs. These developments facilitate the diffusion of advanced data management solutions and analytics. Today, any firm has the opportunity to perform advanced data analytics, which was the privilege of only a few large firms just five years ago. The scale of available services for firms is constantly growing and range from infrastructure platforms to advanced analytics solutions.

Table 1

Techniques and technologies used with Big Data

Techniques for analyzing Big Data	Big Data technologies
– A/B testing	– Business intelligence (BI)
– Classification	– Data warehousing
– Cluster analysis	– Extract, transform, and
– Data mining	load (ETL)
– Machine learning	– Hadoop
– Natural language	– MapReduce
processing (NLP)	– SQL (Structured Query
– Neural networks	Language)
– Network analysis	
– Optimization	
– Predictive modeling	
– Spatial analysis	
– Simulation	
– Time series analysis	
– Visualization	

ETLA

An example of Facebook's Big Data analysis process

Phase 1 – Identify data sources
Gain access to data

(existing historical data and new data)

Phase 2 – Data capture
Collect raw data at servers and prepare the data for storage

Phase 3 – Raw data storage
Store unstructured data in distributed file systems (e.g., Hadoop) (the data can now be queried to create new aggregated datasets)

(Optional – Ad Hoc Querying)
This is an option for real-time and/or complex queries and analysis (typically used for monitoring and complex problem solving)

Phase 4 – Aggregate data storage
Structured format data are stored in traditional data management solutions (typically as subsets of a master database)

Phase 5 – Analysis and reporting
Data analysis from the aggregated structured relational data sets (Statistical analysis and metrics reporting)

What's wrong with current data solutions?

Current data management solutions are typically built around expensive infrastructures, in which data management is centralized and data are accessed in small patches for specific analysis, which often relies on sampling partial data instead of using the entire data set. This is a slow, costly, and inefficient process, particularly if frequent analytic iterations are required – and this is the best-case scenario. The data are often buried within separate organizational silos inside the firm, which can make accessing critical business information difficult and time-consuming. Broadly speaking, the challenge of traditional data systems includes data capture, storage, search, sharing, analysis, and visualization. Because of these obstacles, many firms are not able to utilize one of their most important assets, information, to its fullest potential.

Is it all about data and new technologies?

The Big Data techniques and technologies make the analysis of complex unstructured data feasible and require changes in information technologies, data architectures, and strategic and operative decision-making processes. To take full advantage of the new data opportunities, firms must become data-driven and rely more on advanced data analytics. The essence of this change is to move away from basic business reporting toward using statistical evidence as the basis for decision making. The underlying technologies facilitate the use of data in new ways, but the value and unique opportunities come from data-orientated business thinking. The true challenge is to identify what questions to ask the data.

Who needs Big Data?

Big Data evangelists believe that every business should implement it; however, for most firms, the change is less necessary. The debate about the impact of data-driven business practices is ongoing, but the usefulness of Big Data has already materialized for certain firms. In recent studies, it has been found that when large firms in the U.S. implement Big Data solutions throughout their operations, analytics-based decision making increases productivity and profit by 5–6% compared to firms that have not embraced data-driven decision making.[2]

To identify the sectors with the most potential for gaining value from embracing Big Data, we first examine the amount and type of data in various sectors (Figure 1).

We can see significant differences across industries in the types and volumes of data. Notably,

Figure 1   Types of data generated and stored in various sectors
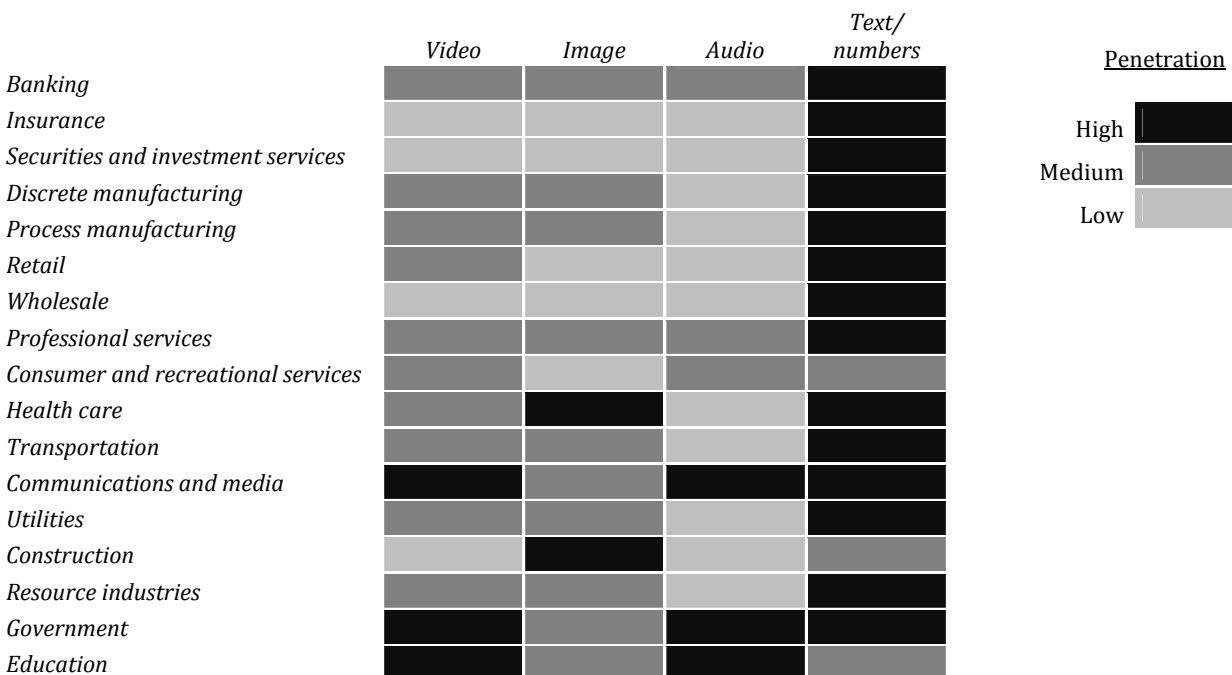
Source: McKinsey Global Institute (2011).

the most data-intensive industries (in which data intensity is high in more than one type of data) are in the public sector (Health care, Government, and Education) and in one private-sector industry (Communications and media). The majority of data are found in the typical more-structured data formats (text and numbers), but unstructured data (audio, image, and video) are abundant, particularly in the public sector industries. In addition, it is estimated that the annual growth rate of data in most sectors is approximately 35% (McKinsey, 2011). Thus, even sectors with low data intensity will likely accumulate a significant amount of data in the near future.

From data intensity, we turn to the business opportunities found in different sectors. We discuss two dimensions, the potential value of data (e.g., cost reductions or value increases) and the ease of capturing value (i.e., level of existing ICT use), using Figure 2 as our starting point.

The most likely first-adopter sectors (because of the data's high value and ease of capture) include the following: Computer and electronic products, Information technology, Transporta-

tion and warehousing, Finance and insurance, and Health care and social assistance. The most likely late-adopter sectors (because of the data's low value and difficulty of capture) include the following: Accommodations and food services; Arts, entertainment, and recreation; and Educational services.

How can firms approach Big Data?

The current trend in most firms that engage Big Data is to use advanced data solutions and traditional data management systems side-by-side in the early stages of adoption. The use of Big Data solutions becomes important when firms begin to expand their existing datasets by creating processes that accumulate larger datasets composed of internal and external data. The challenge for the traditional data systems arises when data analytics are required in real-time with frequent iteration and/or with unstructured data. The most advanced Big Data solutions are able to provide near real-time information solutions through parallel computing that enables large data sets from different sources to be combined and analyzed in multi-patch-based solutions, in addition to analytical tools for reporting and visualization.

Figure 2   **The potential value of data and ease of capture**
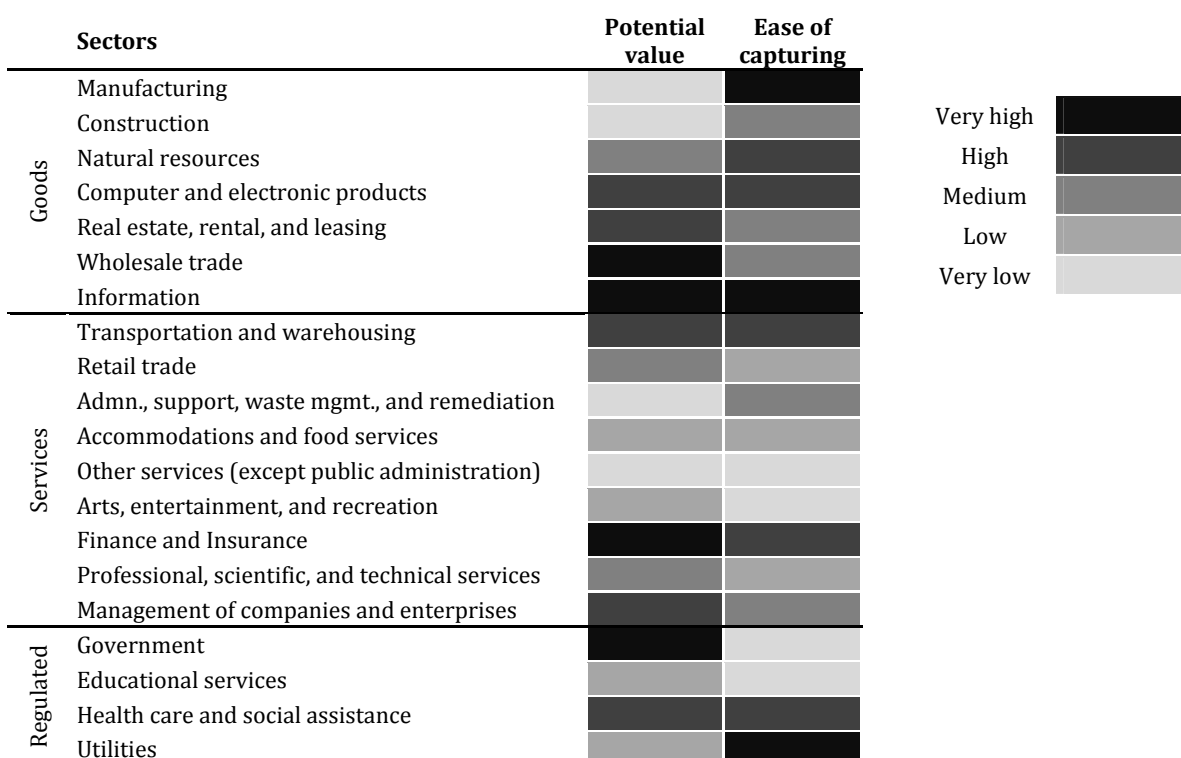
Source: McKinsey Global Institute (2011).

Firms also must choose whether to perform the analytics and data management in-house or outsource these functions. Setting up in-house operations is a viable option for large firms that have ample resources, and a clear need for data-driven strategies. Creating in-house Big Data capabilities requires individuals to be recruited who are in high demand and are difficult to find. Big Data professionals are a combination of statisticians, programmers and industry experts. Finding even one person with these qualities is a challenging feat for any firm; building an entire team with this skill set is a daunting task.

In firms that have built Big Data units, the analytics team is usually organized to operate as internal consultants that tackle specific business unit requirements. This works well in large firms with abundant resources and a constant need for new data insights; however, for many firms with fewer resources and less frequent analytical requirements, the optimal choice is to outsource data management and analytics. If Big Data is outsourced, firms typically retain in-house experts to convey the internal analytics problems to the external partners and to translate the results into concrete business solutions in the firm.

ICT-related outsourcing is a difficult process[3]; for data management and analytics, the challenges might be even greater. The opportunities related to data-driven decision making might only be achieved if there is organization-wide change; when part of the data expertise is outside the company, this change is particularly challenging because integrating external partners into internal process development is notoriously risky. The benefits and disadvantages of building an in-house team compared to using outsourced services is yet to be determined because the number of firms embracing data-driven business practices is only now beginning to grow significantly.

What is the potential for Big Data in Finland?

We have chosen to leave the discussion of the producer side of Big Data technologies outside the scope of this brief, but it requires a few remarks in the context of Finland.

Big Data technology developers are a mix of old and new firms that operate in different layers of the Big Data ecosystem, such as infrastructure (e.g., Amazon), platforms (e.g., Cloudera), and services (e.g., Salesforce).[4] There is significant amount of research and development activity around Big Data, particularly in the services part of the ecosystem. In the near future, many of these Big Data providers (particularly the small ones) will disappear because of consolidation on the provider side as big firms acquire the best technologies to cater to the needs of their established client bases.

Finnish ICT companies most likely will not play a significant role in technology development. The infrastructure and platforms will come from elsewhere (mostly from the U.S.). However, there will be several indirect opportunities as firms begin to set up Big Data activities. Firms have already begun providing analytics and deployment services in Finland (e.g., Tieto), and smaller companies are competing with the bigger domestic actors in niche areas with specialized services.

Provider-side activities are just beginning to emerge. Thus, it is difficult to say how the Big Data infrastructure, platform and service provider landscape will evolve in Finland and elsewhere. However, as the data market grows, it is certain that large international actors will enter the Finnish market, as has occurred in the past with other ICT services.

The user side in Finland will follow the path of many other countries. Few Finnish firms are currently building in-house capabilities in Big Data. By examining recent job openings related to Big Data, we can identify several early adopters (Table 2).

We can easily assume that many more Finnish firms are exploring Big Data opportunities either by building in-house competence or collaborating with external partners.

Table 2	Examples of Big Data users in Finland
Firm	**Industry**
Comptel	Telecom software
Nokia	Mobile phones
Rovio	Games
Sanoma	Media
Supercell	Games
Tieto	Software services

ETLA

Finnish companies and public sector entities are becoming aware of the opportunities that Big Data offers to maintain and/or increase competitive advantages and to create more opportunities to provide high-value-added data-driven services, such as embedded software, which already plays a significant role in several Finnish manufacturing industries.[5] The increasing role of Big Data is not limited to manufacturing; the fastest growing sector in Finland is services, private and public, low and high value-adding.[6] Many of these service sectors create large amounts of data that might be harnessed for decision making in the pursuit of higher productivity.

Is there a need for public policy regarding Big Data?

Much of the policy debate revolves around data privacy and security. There are significant differences in these two aspects around the world.

In the U.S., the landscape is fairly unrestrictive, and many Big Data firms refer to the U.S. as the "Wild West" of data privacy and security. This has allowed U.S. firms to lead the way in Big Data adoption and technology development. However, privacy concerns have reached policy makers in the U.S., who are becoming alarmed by the amount of information that private companies collect on individuals. Thus, investigations on data privacy began to address the scale and scope of data collection in early 2013.

In Europe, these themes have also been topical, and many countries have strong laws regarding data privacy. These strict policies have in part enabled the U.S. to become the leader in the development of Big Data technologies.

Another policy aspect concerns the opportunities for using Big Data in the public sector, where even minor efficiency gains may yield significant social benefits in monetary and social terms. Both public administration and public healthcare might significantly profit from data-driven decision making that significantly reduced costs and increased productivity and general welfare. However, both sectors are notoriously rigid in their approaches to adopting new technologies, and implementing changes to exploit Big Data will likely be as challenging

as it has been to implement changes to accommodate other technologies that have sought to change the public sector for the better.

The final policy remark concerns granting access to publically collected data for business purposes. The public sector collects and stores huge amounts of data that has potentially enormous monetary value that the private sector could harness more efficiently. In several countries, this topic has been debated in recent years, including Finland.[7]

Conclusion

Big Data is here and growing. Whether different sectors, both private and public, can seize the moment and benefit from the opportunities these new solutions provide is a question for future research. The main challenge for engaging Big Data is not technological; instead, the organizational level changes required to take full advantage of analytics-based findings are the most stubborn obstacle. This change is a significant effort for many organizations, particularly in the public sector.

Finland will not be the leader in developing Big Data-related technologies and techniques. However, with the increasing use of Big Data, many ICT sector firms are able to provide Big Data-related services to public- and private-sector users. On the user side, the story might be different. Big Data is already making a significant impact on certain sectors. For example, the digital games industry is among the first adopters of Big Data, and this industry embraced data-driven decision making some time ago. With global competition increasing in almost all sectors, data-driven business practices might provide a solution to maintain and create competitive advantages for Finnish firms. However, the window of opportunity to create competitive advantages through data-driven decision making is slowly closing as more firms adopt data analytics as part of their everyday business. The lead times are already shortening in some of the first-adopter industries, such as the digital games industry, in which most firms and even start-ups are already using advanced analytics to increase revenues.

We are witnessing a significant phase of the ICT revolution in which data are becoming a

valuable asset and a tradable commodity. Simultaneously, we have only limited knowledge about how Big Data and advanced analytics will transform existing industries and spawn new fields. Ongoing research efforts have aimed to shed light on the uses, and potential misuses, of Big Data, but many aspects remain to be explored.

Endnotes

[1]   McKinsey Global Institute (2011). "Big data: The next frontier for innovation, competition and productivity". June 2011.

[2]   Brynjolfsson, E. – Hitt, L. M. – Kim, H.H. (2011). Strength in Numbers: How does data-driven decision-making affect firm performance? Working Paper Available at SSRN: http://ssrn.com/abstract=1819486.

[3]   Koski, H. (2013). ICT outsourcing, user-driven and open innovation strategies in the generation of new data-based solution. ETLA Working Papers no. 7.

[4]   For more on this, see Kushida, K. E. – Murray, J., – Zysman, J. (2011). Diffusing the cloud: Cloud computing and implications for public policy. Journal of Industry, Competition and Trade, 11(3), 209–237.

[5]   Nikulainen, Tuomo – Ali-Yrkkö, Jyrki – Seppälä, Timo (2011). Softaa koneisiin! Ohjelmisto-osaaminen suomalaisen teollisuuden uudistajana. Teknologiateollisuus Ry.

[6]   Pajarinen, Mika – Rouvinen, Petri – Ylä-Anttila, Pekka (2012). Uutta arvoa palveluista. ETLA B 256.

[7]   For more on this discussion in Finland, see Koski, Heli – Kiuru, Pertti – Mäkelä, Jaana – Salokannel, Marjut (2012). Julkinen tieto käyttöön. ETLA Discussion Papers no. 1276.