1

Introduction to the Big Data Era

Stephan Kudyba and Matthew Kwatinetz

CONTENTS

By now you've heard the phrase "big data" a hundred times and it's intrigued you, scared you, or even bothered you. Whatever your feeling is, one thing that remains a source of interest in the new data age is a clear understanding of just what is meant by the concept and what it means for the realm of commerce. Big data, terabytes of data, mountains of data, no matter how you would like to describe it, there is an ongoing data explosion transpiring all around us that makes previous creations, collections, and storage of data merely trivial. Generally the concept of big data refers

to the sources, variety, velocities, and volumes of this vast resource. Over the next few pages we will describe the meaning of these areas to provide a clearer understanding of the new data age.

The introduction of faster computer processing through Pentium technology in conjunction with enhanced storage capabilities introduced back in the early 1990s helped promote the beginning of the information economy, which made computers faster, better able to run state-of-the-art software devices, and store and analyze vast amounts of data (Kudyba, 2002). The creation, transmitting, processing, and storage capacities of today's enhanced computers, sensors, handheld devices, tablets, and the like, provide the platform for the next stage of the information age. These super electronic devices have the capabilities to run numerous applications, communicate across multiple platforms, and generate, process, and store unimaginable amounts of data. So if you were under the impression that big data was just a function of e-commerce (website) activity, think again. That's only part of the very large and growing pie.

When speaking of big data, one must consider the source of data. This involves the technologies that exist today and the industry applications that are facilitated by them. These industry applications are prevalent across the realm of commerce and continue to proliferate in countless activities:

- Marketing and advertising (online activities, text messaging, social media, new metrics in measuring ad spend and effectiveness, etc.)
- Healthcare (machines that provide treatment to patients, electronic health records (EHRs), digital images, wireless medical devices)
- Transportation (GPS activities)
- Energy (residential and commercial usage metrics)
- Retail (measuring foot traffic patterns at malls, demographics analysis)
- Sensors imbedded in products across industry sectors tracking usage

These are just a few examples of how industries are becoming more data intensive.

DESCRIPTION OF BIG DATA

The source and variety of big data involves new technologies that create, communicate, or are involved with data-generating activities, which produce

different types/formats of data resources. The data we are referring to isn't just numbers that depict amounts, or performance indicators or scale. Data also includes less structured forms, such as the following elements:

- Website links
- Emails
- Twitter responses
- Product reviews
- Pictures/images
- Written text on various platforms

What big data entails is structured and unstructured data that correspond to various activities. Structured data entails data that is categorized and stored in a file according to a particular format description, where unstructured data is free-form text that takes on a number of types, such as those listed above. The cell phones of yesteryear have evolved into smartphones capable of texting, surfing, phoning, and playing a host of software-based applications. All the activities conducted on these phones (every time you respond to a friend, respond to an ad, play a game, use an app, conduct a search) generates a traceable data asset. Computers and tablets connected to Internet-related platforms (social media, website activities, advertising via video platform) all generate data. Scanning technologies that read energy consumption, healthcare-related elements, traffic activity, etc., create data. And finally, good old traditional platforms such as spreadsheets, tables, and decision support platforms still play a role as well.

The next concept to consider when merely attempting to understand the big data age refers to velocities of data, where velocity entails how quickly data is being generated, communicated, and stored. Back in the beginning of the information economy (e.g., mid-1990s), the phrase "real time" was often used to refer to almost instantaneous tracking, updating, or some activities revolving around timely processing of data. This phrase has taken on a new dimension in today's ultra-fast, wireless world. Where real time was the goal of select industries (financial markets, e-commerce), the phrase has become commonplace in many areas of commerce today:

- Real-time communication with consumers via text, social media, email
- Real-time consumer reaction to events, advertisements via Twitter
- Real-time reading of energy consumption of residential households
- Real-time tracking of visitors on a website

Real time involves high-velocity or fast-moving data and fast generation of data that results in vast volumes of the asset. Non-real-time data or sources of more slowly moving data activities also prevail today, where the volumes of data generated refer to the storage and use of more historic data resources that continue to provide value. Non-real time refers to measuring events and time-related processes and operations that are stored in a repository:

- Consumer response to brand advertising
- Sales trends
- Generation of demographic profiles

As was mentioned above, velocity of data directly relates to volumes of data, where some real-time data quickly generate a massive amount in a very short time. When putting an amount on volume, the following statistic explains the recent state of affairs: as of 2012, about 2.5 exabytes of data is created each day. A petabyte of data is 1 quadrillion bytes, which is the equivalent of about 20 million file cabinets' worth of text, and an exabyte is 1000 times that amount. The volume comes from both new data variables and the amount of data records in those variables.

The ultimate result is more data that can provide the building blocks to information generation through analytics. These data sources come in a variety of types that are structured and unstructured that need to be managed to provide decision support for strategists of all walks (McAfee and Brynjolfsson, 2012).

BUILDING BLOCKS TO DECISION SUPPORT

You may ask: Why are there classifications of data? Isn't data simply data? One of the reasons involves the activities required to manage and analyze the resources that are involved in generating value from it. Yes, big data sounds impressive and almost implies that value exists simply in storing it. The reality is, however, that unless data can help decision makers make better decisions, enhance strategic initiatives, help marketers more effectively communicate with consumers, enable healthcare providers to better allocate resources to enhance the treatment and outcomes of their patients, etc., there is little value to this resource, even if it is called big.

Data itself is a record of an event or a transaction:

A purchase of a product
A response to a marketing initiative
A text sent to another individual
A click on a link

In its crude form, data provides little value. However, if data is corrected for errors, aggregated, normalized, calculated, or categorized, its value grows dramatically. In other words, data are the building blocks to information, and information is a vital input to knowledge generation for decision makers (Davenport and Prusak, 2000). Taking this into consideration, the "big" part of big data can actually augment value significantly to those who use it correctly. Ultimately, when data is managed correctly, it provides a vital input for decision makers across industry sectors to make better decisions.

So why does big data imply a significant increase in the value of data? Because big data can provide more descriptive information as to why something has happened:

Why and who responded to my online marketing initiative?
What do people think of my product and potentially why?
What factors are affecting my performance metrics?
Why did my sales increase notably last month?
What led my patient treatment outcomes to improve?

SOURCE OF MORE DESCRIPTIVE VARIABLES

Big data implies not just more records/elements of data, but more data variables and new data variables that possibly describe reasons why actions occur. When performing analytics and constructing models that utilize data to describe processes, an inherent limitation is that the analyst simply doesn't have all the pertinent data that accounts for all the explanatory variance of that process. The resulting analytic report may be missing some very important information. If you're attempting to better understand where to locate your new retail outlet in a mall and you don't have detailed shopper traffic patterns, you may be missing some essential