Privacy in the Age of Big Data: Exploring the Role of Modern Identity Management Systems

Ali M. Al-Khouri[1, 2]

[1] Emirates Identity Authority, Abu Dhabi, United Arab Emirates

[2] British Institue of Technology and E-Commerce, London, U.K.

Correspondence: Dr. Ali M. Al-Khouri, Emirates Identity Authority, Abu Dhabi, United Arab Emirates. E-mail: ali.alkhouri@emiratesid.ae

Received: August 7, 2013 Accepted: September 27, 2013 Online Published: September 29, 2013

doi:10.5430/wjss.v1n1p37 URL: http://dx.doi.org/10.5430/wjss.v1n1p37

Abstract

In today's digital world, our ability to better understand data is seen as fundamental to addressing complex economical and societal challenges. The massive amounts of digital data that governments and businesses collect as well as the technological tools they use for analyzing disparate data are referred to as big data. These advances in data collection and analysis have raised concerns about individuals' rights to privacy. In this article, we attempt to provide a short overview of big data and explore the role of modern identity management systems in providing higher levels of security and privacy in online environments. The article also makes reference to one of the most advanced identity management systems in the world, namely the United Arab Emirates' (UAE) identity management infrastructure, and how the government has designed its systems to support privacy and security in e-government and e-commerce scenarios.

Keywords: big data, privacy, identity management

1. Introduction: The Age of Big Data

The proliferation of modern technologies and smart devices in addition to the popularity of social networking is generating unprecedented amounts of data, both in structured and unstructured forms, whether it be text, audio, video, or other forms (Mayer-Schonberger and Cukier, 2013). Data as a term and concept has become ubiquitous in today's digital landscape. In fact, data is rather becoming multi-form, multi-source, and multi-scale (Sathi, 2013).

The sheer number of bytes that is generated daily is mind-boggling! According to a recent report published by IBM, 2.5 quintillion (2.5×10^{18}) bytes of data are created every day (IBM, 2010). According to the same report, 90% of the data in the world today has been created during the past two years. Experts indicate that the world is in a digital explosion era (Liebowitz, 2013; Marz and Warren, 2013; Minelli et al., 2013; Smolan and Erwitt, 2012). This phenomenon is referred to as *big data*.

"Big data" refers to a conglomerate of datasets whose size is beyond typical database software's ability to capture, store, manage, and analyze (Manyika et al., 2011). As depicted in Figure 1, all computer hard drives in the world equaled 160 exabytes in 2006. The total storage systems did not reach one zetabyte of information in 2012. One Zetabyte equals to 1,000,000,000,000,000,000,000 bytes, or 1000 exabytes. By 2020, the expected growth rate is forecasted to reach 112 zetabytes of data, representing almost 75% annual growth rate.

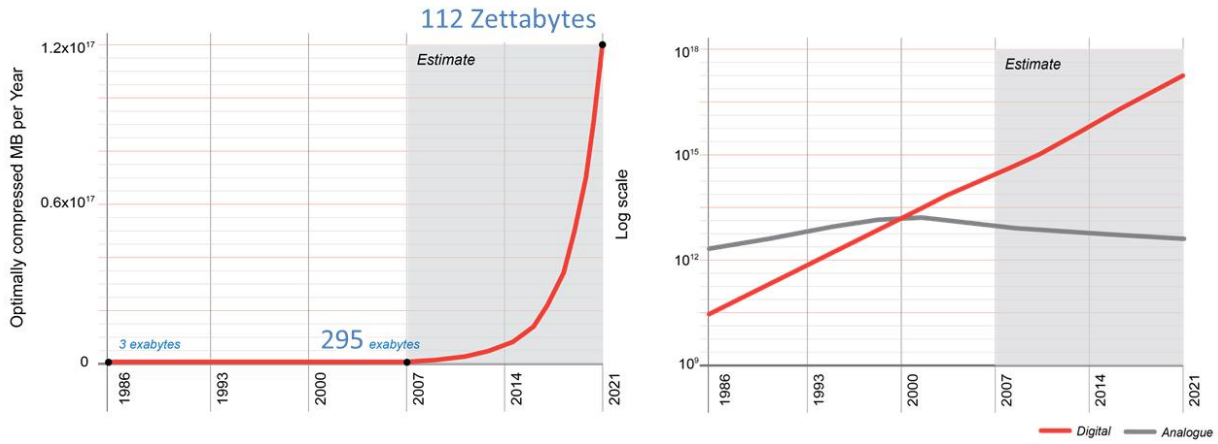

Figure 1. Global Growth of Digital Storage Capacity

Big data is generated from practically everywhere; i.e., social media sites (Facebook, Twitter, Linked-in), digital pictures and videos, e-mails, purchase transaction records, cell phone, global positioning system (GPS) signals, geo-stationary satellites, and meteorological sensors, to name a few. See also Figure 2. Billions of posts in social networks, blogs, commerce sites, e-mails, text messages, and utility payments are being "piggy-backed" to result in patterns of the digital interactions and individual behavior patterns that are then constructed from there.

Figure 2. Big Data Sources

Data growth is being enabled by innovative software and analysis tools as well as inexpensive storage and a proliferation of sensor and data capture technology, thus increasing connections to information via the cloud and virtualized storage infrastructures (Gantz and Reinsel, 2011). A study that IDC conducted in 2011 showed that new technologies are driving down the cost of creating, capturing, managing, and storing information to one-sixth of the cost in 2005 (ibid.). See also Figure 3. The same study also indicated that consumers will create almost 68% of unstructured data in 2015.

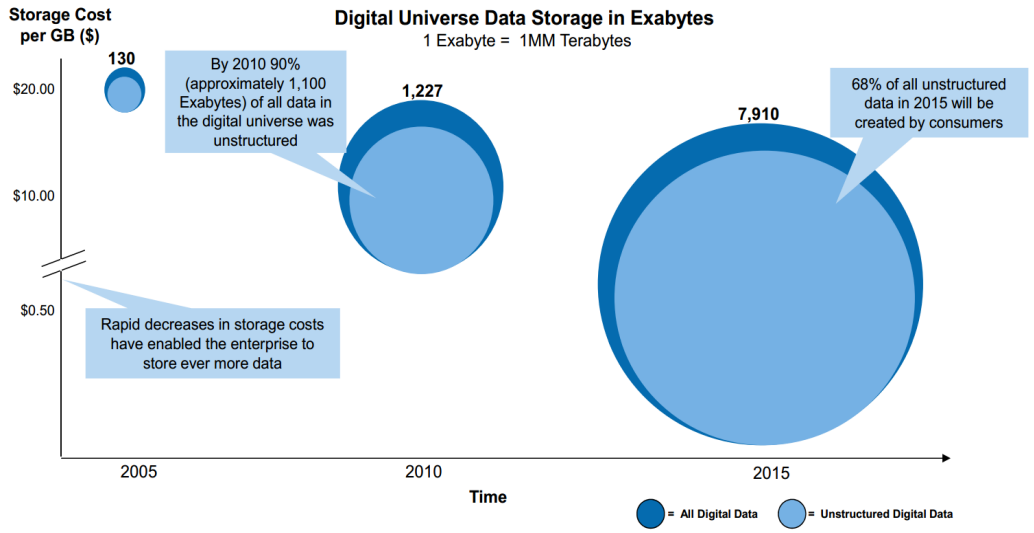

Figure 3. The Growth of Unstructured Data. Source (IDC, 2011)

All in all, the amount of data is continuing to grow at an exponential rate. As it grows, this collection of data is seen as creating a new layer in the economy by turning information into revenue and accelerating growth in the global economy by creating jobs (Gartner, 2012).

This article's purpose is to explain this phenomenon and to view it from an individual's privacy perspective. The article mainly attempts to shed light on modern identity management systems' role in protecting individuals' privacy rights in online environments. We use the example of the United Arab Emirates (UAE) and its identity management infrastructure in this regard.

The article is structured as follows. In section 2, we explain the characteristics that constitute big data. In section 3, we provide some thoughts regarding how identities can be constructed from online and digital behaviors. In section 4, we illustrate how government identity management systems can provide higher security and protection levels in online environments. We also demonstrate how the UAE's government has addressed the privacy and security concerns of its citizens and residents in online e-government and e-commerce transactions. The article is then concluded in section 5.

2. Big Data Characteristics

Big data has come to be characterized by the *volume, velocity, and variety* of data that is generated. These constitute the 3Vs of big data. See Figure 4. Volume refers to the amount of data and the form of data. Velocity refers to the rate at which the data are collected and analyzed. Meanwhile, variety provides the type of data collected.

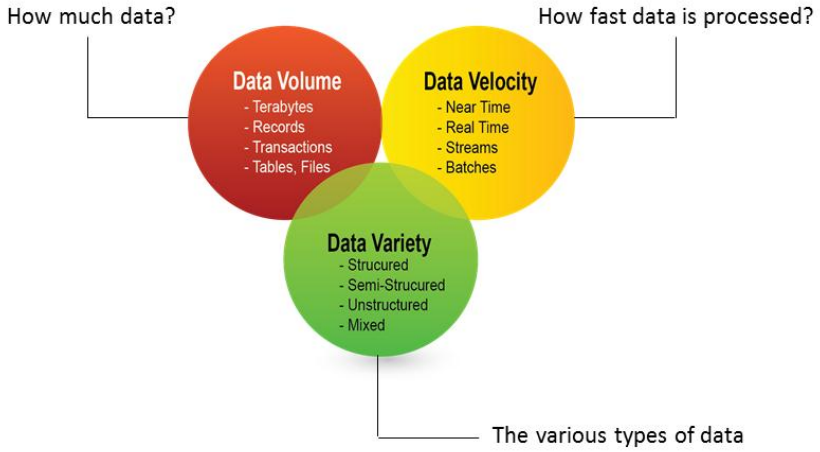

Figure 4. Big Data Characteristics. Source: Russom (2011)

More recently, big data has been characterized by an additional fourth dimension—a fourth V—veracity, which encompasses the 3Vs. Veracity provides confidence in the truthfulness of the data. Veracity of data itself can be depicted using three dimensions. See also Figure 5.Veracity of data is established by how the data itself is enabled, which stands for the source of data. Veracity of data is established by the means and methods of analysis, thus providing discernible information. Veracity of data is then characterized by personal identity management to impact business outcomes. This is the critical dimension of big data veracity.

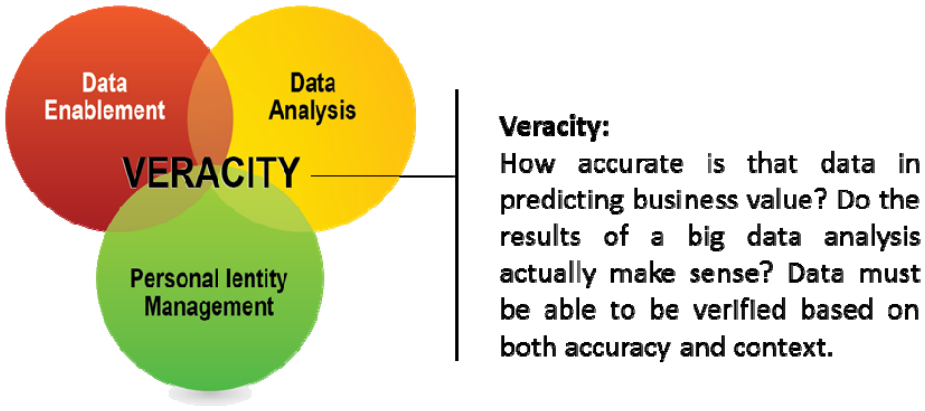

Figure 5. The 3 Dimensions of Veracity

In principle, big data is not a new or unknown phenomenon. In fact, big data as volume data has been used in clinical trials for a long time, resulting in many groundbreaking innovations in medicines, for example. In addition, big data as volume data has been in existence in deoxyribonucleic acid (DNA) mapping and genetics, leading to many life-saving healthcare procedures. While the healthcare industry has been the initiator of big data analysis, retailers and marketing organizations have now begun to make use of big data to further their commercial activities.

Overall, the use of big data varies across sectors, where some sectors are poised for greater gains. Figure 6 depicts the results of an analysis that McKinsey conducted in 2011 and illustrates differences among sectors in the use of big data (Manyika et al., 2011). The study divided the sectors into primarily 5 clusters. These include Cluster A: computer and electronic products; Cluster B: finance, insurance and government; Cluster C: construction, educational services, and arts and entertainment; Cluster D: manufacturing and wholesale trade; and Cluster E: retail, health care providers, accommodation, and food.

Cluster A sectors have already posted very strong productivity growth and are set to gain substantially from using big data since they have access to huge pools of data, and the pace of innovation is very high. Cluster B sectors, which include finance, insurance, and government, are positioned to benefit very strongly from big data as long as barriers to its use can be overcome. Because both clusters A and B are transaction- and customer-intensive sectors, the use of data and experimentation is envisaged to drastically improve overall performance. Clusters C, D, and E can derive significant value from big data, although doing so will depend on the extent to which barriers are overcome.

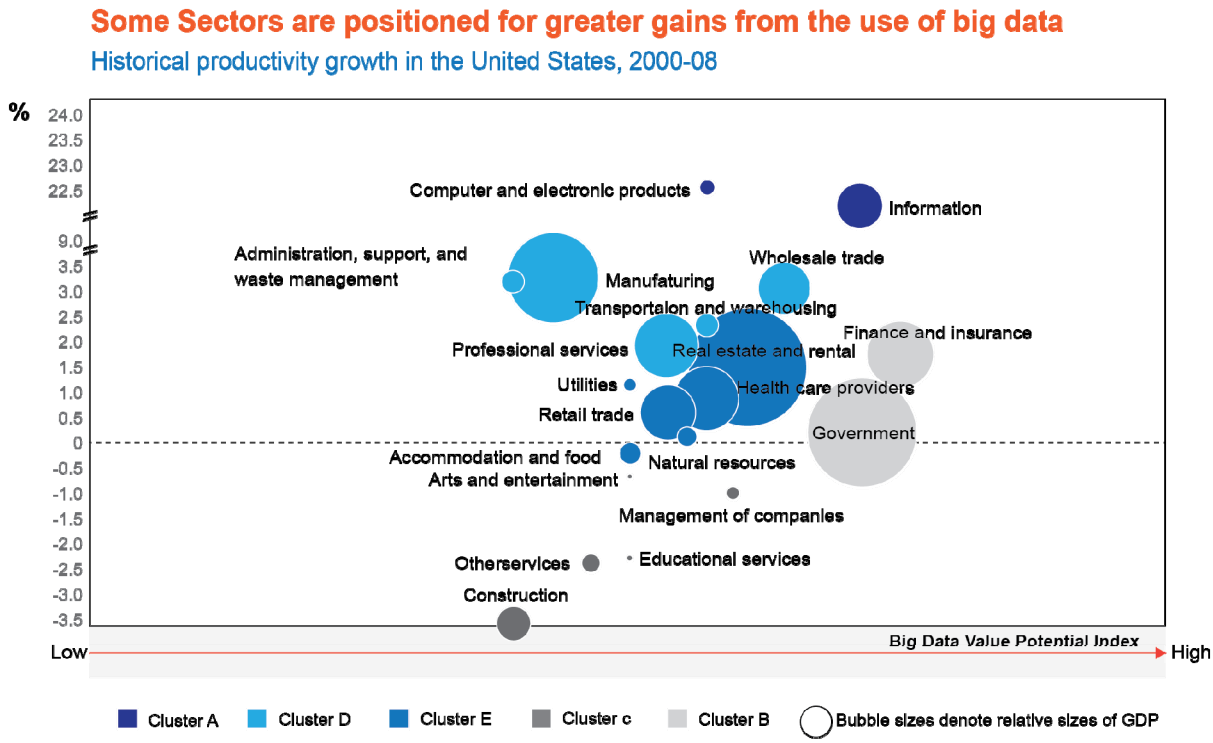

Figure 6. Big Data Value across Sectors. Source: (Manyika et al., 2011)

According to Gartner, "data-driven innovation," will help to create 4.4 million information technology (IT) jobs globally by 2015, including 1.9 million in the United States (US) (Gartner, 2012). McKinsey's report indicates that big data has the potential to create massive saving and revenues in all sectors, i.e., create $300 billion in potential annual value to U.S. health care (more than double the total annual health care spending in Spain); €250 billion potential annual value to Europe's public sector administration (more than the gross domestic product [GDP] of Greece); and $600 billion in potential annual consumer surplus from using personal location data globally (Manyika et al., 2011).

All in all, big data is considered to have a huge impact on all sectors, providing endless arrays of new opportunities for transforming decision-making; discovering new insights; optimizing businesses; and, innovating their industries. However, with all of this data out there in the hands of "others," how can privacy be achieved for the individual? In fact, this could be construed as a blatant violation of individual privacy. Let us explore this further in the next section.

3. Constructing Identity from Digital Behaviour

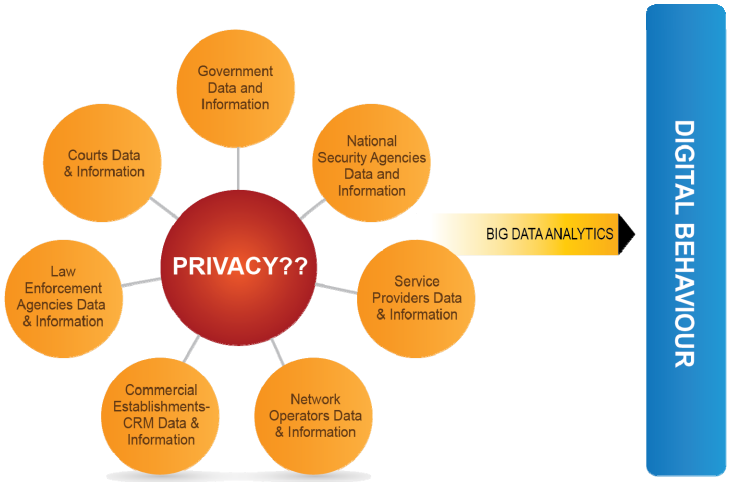

Figure 7. Digital Behavior and Privacy

Big data in information and datasets is captured based on the digital data we leave in our communications and transactions. See also Figure 7. In every interaction, we leave behind a huge trail of data that includes bits and pieces and pointers to our real behaviors. For reasons explained in earlier sections, we seek to analyze our communications using the argument that entities need to know their customers in order to predict our preferences and to enable the personalization of services and products based on our needs. This, in turn, raises many issues that govern privacy and our rights to be anonymous in the digital world.

In principle, it is understood that the collection of information from digital transactions and interactions is something that is unstoppable. Whether we like it or not, the digital trail we leave behind in the e-world is amazingly large. This digital trail when analyzed is almost like a signature that we leave behind, making it very easy for analysts to identify us as individuals in the purported anonymity of the World Wide Web (WEB).

O'Harrow (2006) indicates that although the emergence of a data-driven surveillance society has provided the conveniences of access to information and services (such as cell phones, discount cards, and electronic toll passes), it also has created new approaches to watching us more closely than ever before. He also points to the fact that as companies customarily collect billions of details about nearly every connected individual, the world will reach a state where people will lose control of their privacy and identities at any moment. Figure 8 depicts an illustrative diagram of the evolving possibilities of capturing a data trail of individuals in the digital world.

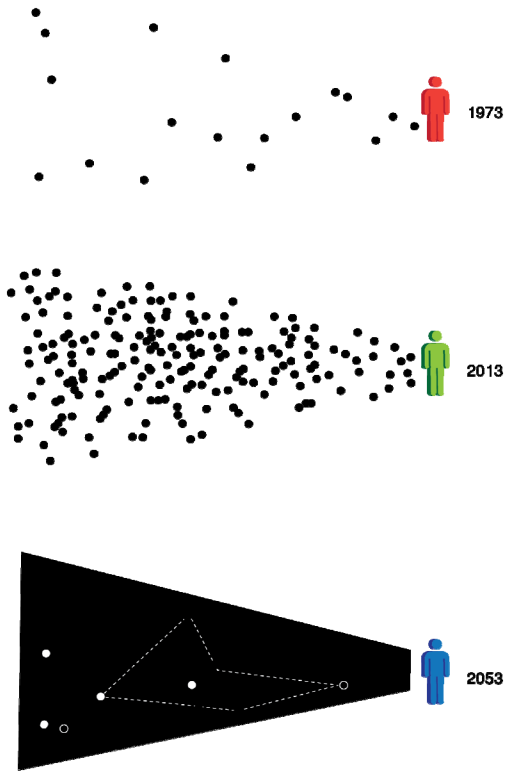

Figure 8. Digital Behavior and Data Trail in Big Data

On a global perspective, the West, particularly the US and the European Union (EU), have made conscious moves to protect individuals' privacy from being abused using legal provisions. Anonymity has been the key consideration on which the legal provisions have been made so far. However, it has been proved beyond any reasonable doubt that anonymity is not guaranteed even when personal identifiers are removed from the data sets for analysis. Personal information can be revealed through searches by the user's computer, account, or Internet protocol (IP) address being linked to the search terms used (Blakeman, 2010). Thus, where does this leave an individual with respect to privacy?

Ohm (2009) says that possibilities always exist to re-identify or de-anonymize the people hidden in an anonymized database and that "data can be either useful or perfectly anonymous but never both." In addition, Masiello and

Whitten (2010) indicate that even anonymized information will always carry some risk of re-identification:

> ".... many of the most pressing privacy risks... exist only if there is certainty in re-identification… that is if the information can be authenticated. As uncertainty is introduced into the re-identification equation, we cannot know that the information truly corresponds to a particular individual; it becomes more anonymous as larger amounts of uncertainty are introduced." (p. 122)

Masiello and Whitten (2010) also indicate that a need exists for the development of not just a set of principles and policies but also a set of technical solutions that give users meaningful control. The next section attempts to present how modern identity management systems can address this need, i.e., privacy protection.

4. Government Identity and Privacy Constructing

Many governments throughout the world have launched modern identity management systems, aiming in principle to strengthen national security (Al-Khouri, 2012). Such systems attempt to establish unique identifications of individuals and to provide government-issued personal identity cards and digital identity profiles.

Digital identity profiles provide perfect PROXY for personal identities. Individuals would be known and authenticated as genuine persons by a "national identity authority" that will act as a third-party, online identity authentication service provider. In online transactions, no identity details are revealed to the service providers apart from basic identity details. Thus, service providers, in turn, can identify the potential service-seekers securely from the authentication that the identity authority provides. An individual will then be able to transact and interact freely without compromising his/her personal identity in e-government and e-commerce applications. See also Figure 9.

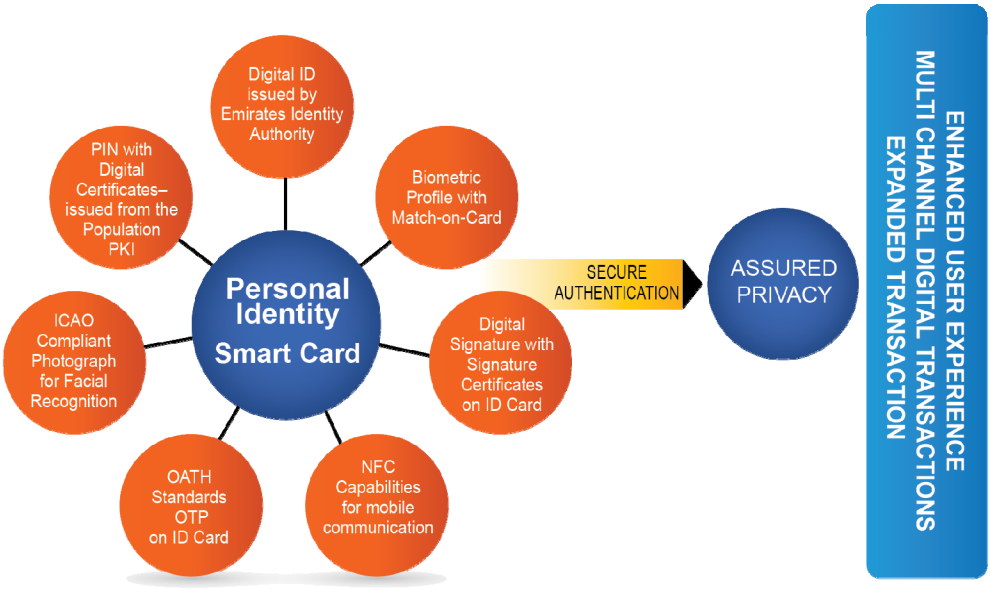

Figure 9. Stronger Personal Identity with Government Identity Systems

From a government perspective, such systems are envisaged to be extremely critical in big data and big data analytics in the sense that they provide the required privacy in anonymity yet provide meaningful data for analysis.

4.1 UAE National Identity Program

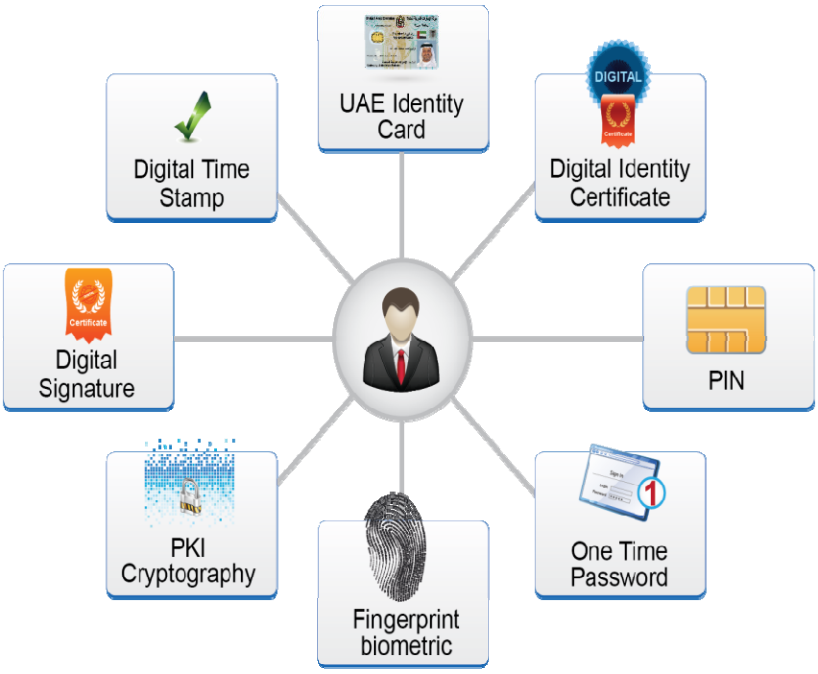

Figure 10. UAE identity card authentication capabilities

The UAE national identity management infrastructure was set up in 2005, and all citizens and residents were registered in the system by the end of 2012. All of the population has been issued smart biometric and Public Key Infrastructure (PKI)-based identity cards, with biometric enrollment being mandatory for those above the age of 15 years.

The smart cards issued are designed to provide multi factor authentication. The digital identity profile components in the card provide the ability to verify and to authenticate the identity of the individual for access.

An online validation gateway has been set up in the UAE to provide the necessary credential verification on the Web. The identity card could only be used with the digital credentials on Web transactions. The validation gateway does not share personal information but provides credential verification. As such, service providers are accorded with verification and authentication services that enable secure remote transactions. Service-seekers remain anonymous on the Web because only digital certificates or biometrics would be used to establish credential verification. See also Figure 11.

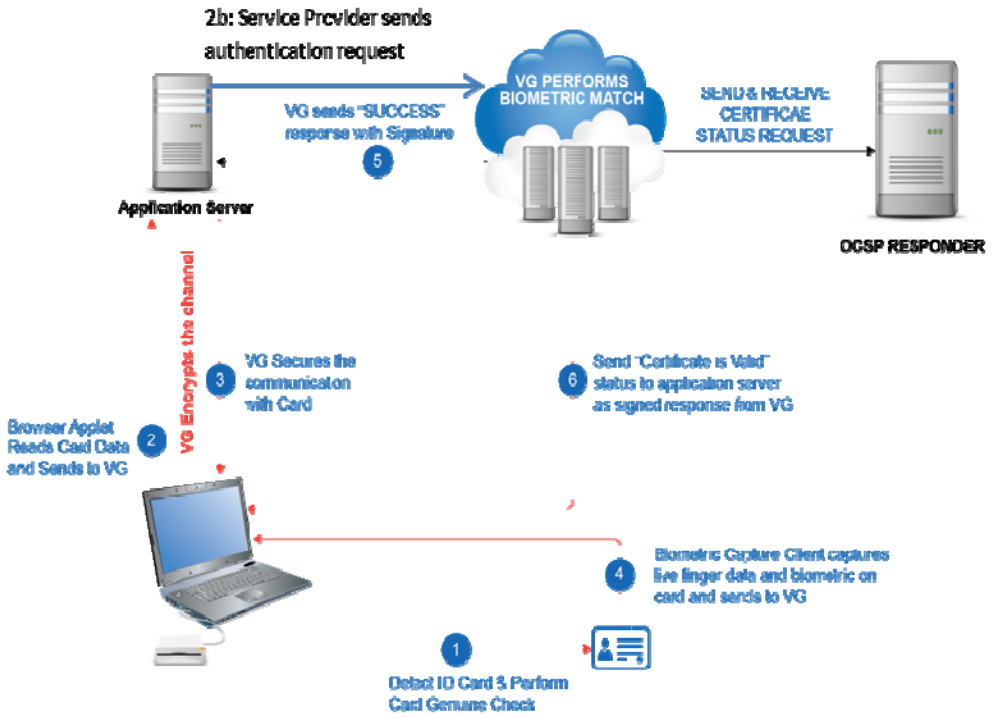

Figure 11. Government Identity Card and Protection of Individual Privacy

Anonymity to the service seeker is assured because no personal details are transmitted across the network channels. The service provider relies on the digital identity credentials provided in the national identity card. When the identity card is presented in the transaction, the service provider simply refers the credentials to the *identity provider* (national identity authority). The identity authority, in turn, verifies the credentials, establishes the credentials' validity, and sends back a digitally signed response that verifies the cardholder's identity.

All of the interactions on identity verification are done using standard protocols of digital certificate verification. The true identity linked to these credentials is only with the *service provider*, the *identity provider*, and the *ID holder*.

The service provider is pleased with the fact that the presence of the correct entity is established in the transaction. The service-seeker is satisfied on the basis that none of his/her personal details are out in the open and that privacy is assured. Snoopers on the transaction collecting digital trails only get bits and bytes of data with no information on them. It is important to note that the information from the transaction remains only with the service-seeker and the service provider.

Let us consider a simple transaction on the Web where a purchase is affected. Let us assume that the seller on the Web has a policy of selling only to people above 18 years of age. Under the current conditions, the buyer online is expected to complete a form with personal details, such as name, address, gender, date of birth, etc., and sign a disclaimer that he/she is above 18. These data are worth their value in gold for snoopers. While the service provider seeks this information to protect his/her selling policies, the service-seeker is forced to provide verifiable information that the snoopers happily gather.

With the advent of the national identification card, the service seekers information is "read" off the card using secure applications, the identity is verified and signed digitally by the identity provider, the age information is verified digitally with no personal data being transferred across the networks. While big data collectors and snoopers can get valid information about a sale indicating that a person above 18 has transacted, the buyer's identity is fully protected and is not divulged on the public channels.

5. Concluding Remarks

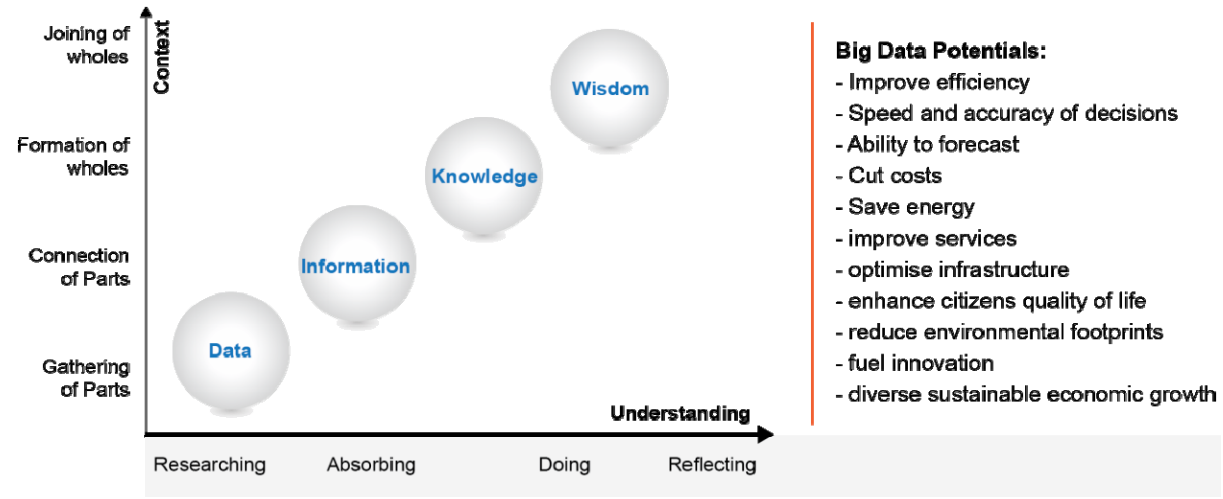

Figure 12. Knowledge-building through Big Data

Despite all fears associated with it, big data should be viewed as being about building knowledge to support social, environmental, and economic development. However, complexity will remain an issue. Successfully exploiting the value in big data requires experimentation and exploration. The private sector will still lead the game, as big data will be viewed as a source of competition and growth.

The public sector will need to take big data more seriously and put in place data strategies to create new waves of productivity growth. The shortage of skills will be a primary challenge. It is reported that the US by 2018 will face a shortage of about 2 million managers and analysts who have the know-how needed to create and use big data to make effective decisions (Manyika et al., 2011).

Conversely, the notion of trust in how information is used, shared, archived, and managed is critical in this complex and highly fluid environment (Gantz and Reinsel, 2011). Governments will need to pay more attention to addressing policies that are related to privacy and security needs in today's digital world. From our perspective, we believe that data in whatever form should be treated as personally identifiable and as a result should be subjected to the regulatory framework.

In this article, we highlighted the potential role of modern government identity management systems in providing higher levels of privacy and security in online transactions. The presented case of the UAE provides a real case of a government practice in this field. Digital identity profiles provided and packaged in secure smart cards can be expected to play a pivotal role in balancing the needs of service providers and service-seekers. A secure identity would encourage users to be engaged more actively and more expansively in this digital world.

Acknowledgments

The content of this article was presented at Big Data Systems, Applications and Privacy Conference, organized by New York University, Abu Dhabi, UAE, 10 –11 March 2013.

References

Al-Khouri, A. M. (2012). Biometrics Technology and the New Economy. *International Journal of Innovation in the Digital Economy*, *3*(4), 1-28. http://dx.doi.org/10.4018/jide.2012100101

Blakeman, K. (2010). What Search Engines Know About You. *Online*, *34*(5), 46-48.

Cooper, M., & Mell, P. (2012). *Tackling Big Data*. National Institute of Standards and Technology. US Department of Commerce. Retrieved from http://csrc.nist.gov/groups/SMA/forum/documents/june2012presentations/fcsm_june2012_cooper_mell.pdf

Craig, T., & Ludloff, M. E. (2011). *Privacy and Big Data*. Sebastopol: O'Reilly Media.

Gantz, J., & Reinsel, D. (2011). *Extracting Value from Chaos*. IDC. Retrieved from http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf

Gartner. (2011). *Big Data Creates Big Jobs: 4.4 Million IT Jobs Globally to Support Big Data By 2015*. Retrieved from http://www.gartner.com/newsroom/id/2207915

Hagen, C., Khan, K., Ciobo, M., Miller, J., Walll, D., Evans, H., & Yadav, A. (2013). *Big Data and the Creative Destruction of Today's Business Models*, A.T. Kearney, Inc. Retrieved from https://www.atkearney.com/documents/10192/698536/Big+Data+and+the+Creative+Destruction+of+Todays+Business+Models.pdf/f05aed38-6c26-431d-8500-d75a2c384919

IBM. (2010). *What is big data? Bringing big data to the enterprise*. Retrieved from http://www-01.ibm.com/software/data/bigdata

Liebowitz, J. (2013). *Big Data and Business Analytics*. Verlag: Auerbach Publications. http://dx.doi.org/10.1201/b14700

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. Retrieved from http://www.mckinsey.com/~/media/McKinsey/dotcom/Insights and pubs/MGI/Research/Technology and Innovation/Big Data/MGI_big_data_full_report.ashx

Marz, N., & Warren, J. (2013). *Big Data: Principles and best practices of scalable realtime data systems*. NY: Manning Publications.

Masiello, B., & Whitten, A. (2010). Engineering privacy in an age of information abundance. In Intelligent Information Privacy Management, AAAI Spring Symposium Series, 119-124. Retrieved from https://www.aaai.org/ocs/index.php/SSS/SSS10/paper/view/1188/1497

Mayer-Schonberger, V., & Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston, MA: Eamon Dolan/Houghton Mifflin Harcourt.

Minelli, M., Chambers, M., & Dhiraj, A. (2013). *Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses*. New Jersey: John Wiley & Sons. http://dx.doi.org/10.1002/9781118562260

Nair, R., & Narayanan, A. (2012). *Benefitting from Big Data Leveraging Unstructured Data Capabilities for Competitive Advantage*. Booz & Company Inc. http://www.booz.com/media/file/BoozCo_Benefitting-from-Big-Data.pdf

O'Harrow, R. Jr. (2006). *No Place to Hide*. New York: Free Press.

Ohm, P. (2010). Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. *UCLA Law Review*, *57*, 1701. Retrieved from http://ssrn.com/abstract=1450006

Russom, P. (2012). *Big data analytics*. Retrieved from http://www.cloudtalk.it/wp-content/uploads/2012/03/1_17959_TDWIBigDataAnalytics.pdf

Sathi, A. (2013). *Big Data Analytics: Disruptive Technologies for Changing the Game*. USA: Mc Press.

Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D., & Tufano, P. (2012). *Analytics: The real-world use of big data: How innovative enterprises extract value from uncertain data*, IBM Corporation, USA. Retrieved from http://www-03.ibm.com/systems/hu/resources/the_real_word_use_of_big_data.pdf

Smolan, R., & Erwitt, J. (2012). *The Human Face of Big Data*. Sausalito, Calif.: Against All Odds Productions.

United Nations. (2012). *Big Data for Development: Challenges and Opportunities*. Global Pulse. Retrived from http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf