Natural Resources Canada
Ressources naturelles Canada

CANADIAN GEOSPATIAL DATA INFRASTRUCTURE
INFORMATION PRODUCT 44e

Impacts and Implications of Big Data for Geomatics: Backgrounder

GeoConnections
Hickling Arthurs Low Corporation

2016

Canada

Impacts and Implications of Big Data for Geomatics: Backgrounder

Prepared for:

Natural Resources Canada
GeoConnections

2014

GeoConnections
GéoConnexions
Canada

Hickling Arthurs Low
Innovation Policy Economics

Acknowledgements

GeoConnections would like to acknowledge the contributions made by Intelli3– Dr. Yvan Bédard and Sonia Rivest, for providing research, writing and editing of the Backgrounder, Hickling Arthurs and Low Corporation – Ed Kennedy, for supervising the production and revision process. We also acknowledge the contribution of Dr. Monica Wachowicz from the University of New Brunswick who made excellent suggestions in her review of the draft Research Report that was produced prior to this GeoConnections Backgrounder. The staff at GeoConnections who provided management, input and direction for the project were Dr. Jean Brodeur, Simon Riopel, Cindy Mitchell and Paula McLeod. In addition Eric Wright, Craig Eby, Alexandre Cyr and Laurent St-Arnaud provided assistance in assembling this document.

Table of Contents

Executive Summary

This backgrounder was written to further examine the characteristics of Big Data and the impact it has with the geomatics sector. The first chapter of the report discusses key concepts of Big Data in detail including its origins, definition, actors and usages. The second chapter of the report addresses the geospatial side of Big Data, starting with the contribution of geomatics to Big Data and vice-versa. The last chapter concludes with key elements to remember, followed by the appendices and references.

Big data has been an emerging trend in geomatics in recent years with datasets continuing to get larger and more complex resulting in new challenges for organizations managing and analyzing data. Geospatial information technologies will now play a central role into how successful and useful Big Data can become in the geomatics sector. The emergence of Big Data brings a need for embedding powerful analytics into organizations that will create additional value from the location base information collected.

Geomatics contributes to the success of Big Data in three main ways: by enriching data visualization, integrating unlinked big data and more powerful analytics. Many large companies such as Facebook, Amazon and Walmart are already currently using Big Data analytics to link user locations with activities, track inventory and other valuable location based information. Big Data also benefits from geospatial solutions in various ways from technologies such as digital maps produced from satellite imagery, aerial photographs and GPS-based field measurements.

Big Data (along with Business Intelligence (BI)) concepts are also impacting the geomatics community in two main ways, as a facilitator, and as a source of innovation. New innovations include new geospatial-specific solutions, new bodies of knowledge, new scientific communities, new specialized conferences, and new working groups in standardization bodies (e.g., OGC). Big Data is enabling communities to perform new tasks such as combining analytics capabilities with the transactional approach of Geographical Information Systems (GIS) to provide new insights in spatially-referenced business data. Big data facilitates the geomatics industry with access to powerful, scalable storage and processing services hosted at remote locations, which further simplifies existing work being done.

The rapid growth of Big Data brings many challenges as well for the geomatics sector. The five main challenges that are currently in need of solutions and strategies include: location privacy, embracing the new paradigm in geomatics, geomatics skills, spatial interoperability, and geospatial data processing technology. Big Data has also created new opportunities for geomatics communities as experts in geomatics know how powerful geospatial data is with regards to integration, analytics and visualization. Big Data has already changed the way location based data can be managed and explored and with new possibilities arising information technologies must be implemented to effectively process and analyze the data.

Preamble

This guide is one in a series of Operational Policy documents being developed by GeoConnections. This document is intended to inform CGDI stakeholders about "Big Data" in terms of technologies, usages, challenges, and opportunities, and more specifically how geospatial data and Big Data are connected.

The use of the term "Big Data" is widespread. Although there is a lot of hype around this concept, it is a major trend creating a real revolution that will impact many aspects of our society. This backgrounder was written to further examine the characteristics of Big Data and the impact it has with the geomatics sector. The first chapter introduces the Big Data concept, including its origins, definition, actors and usages. The second chapter introduces the relationship of Big Data and geomatics, which includes a discussion of the role of geospatial data in the Big Data ecosystem. In addition, risks in terms of legal or policy related concerns, such as privacy, opportunities and the value proposition of the geospatial component 'Big Data' are addressed. The last chapter concludes with key elements to remember, followed by the appendices and references.

The **GeoConnections** *program is a national initiative led by Natural Resources Canada. GeoConnections supports the integration and use of the* **Canadian Geospatial Data Infrastructure (CGDI).**

The **CGDI** *is an on-line resource that improves the sharing, access and use of Canadian geospatial information – information tied to geographic locations in Canada. It helps decision makers from all levels of government, the private sector, non–government organizations and academia make better decisions on social, economic and environmental priorities.*

1. Key Concepts of Big Data

1.1 What is Big Data?

Consensus on a rigorous definition of Big Data has not been reached (Mayer-Schönberger & Cukier, 2013), (Franks, 2012), (McKinsey Global Institute, 2011). While definitions vary (Wehbe, 2013), there seems to be a common understanding, at least in the scientific community, that Big Data simultaneously includes three characteristics. As stated in the definition proposed by Gartner: "***Big Data are 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, 2012). Volume, velocity and variety are called the "3 Vs" of Big Data and are considered their fundamental characteristics. (see Appendix B).

Big Data is exploited by a new generation of technologies designed to extract valuable information from very large volumes of a wide variety of data, by enabling high-velocity capture and analysis. It is assumed that, as technology advances over time, the size of datasets that qualify as Big Data will also increase. The definition can vary by sector, depending on what kinds of software tools are commonly available and what sizes of datasets are common in a particular industry (McKinsey Global Institute, 2011).

Harnessing Big Data allows one to complement traditional data with mass-provided data and new analytics to produce almost real time predictions regarding a phenomenon without necessarily understanding the underlying causes of the behavior of this phenomenon (Mayer-Schönberger & Cukier, 2013). This transition to a data-centric era allows us to work at new scales to extract insights in new ways.

1.2 The Origins of the Big Data Concept

The world is witnessing an unprecedented flood of data and data-related technologies. The term “Big Data” first appeared in the mid-1990s in the scientific communities (e.g., astronomy and genomics) that encountered this data explosion of larger and larger volumes of heterogeneous, fast evolving data being produced and available for analysis. From logs of web activities (e.g., Amazon, Google) to social networks (e.g., Facebook, Twitter) and smartphone (e.g., Apple, Samsung) interactions, data streams from real-time sensors (e.g., Intelligent Transportation Systems (ITS) sensors and surveillance cameras), Radio-Frequency Identification (RFID) tags in supply chain management, to Earth observation satellites, the sources of data are multiplying faster and faster while becoming more and more fine grained, varied and timely.

By 2007 – 2008, the term Big Data started to become familiar via web blogs, marketing material and job postings. 2010 is often considered the Big Data year. This is when the concept became a phenomenon recognized by professionals and was finally accepted in Wikipedia after previous rejections, and showed up in Google search trends (McBurney, 2012). It made the Gartner Hype Cycle for Emerging Technologies in 2011 (Fenn & LeHong, 2011), where it was identified as one of the fastest maturing technologies. Since then, it has been widely spread by the Information Technology (IT) community, especially by companies selling storage hardware, cloud storage and services, data warehousing and analytics solutions. The convergence of two interrelated trends made the Big Data phenomenon possible: (1) increased affordability of technologies (e.g. flash-based memory, large storage, sensors, smartphones, clouds, open-source software), and (2) increased production of data (e.g., social networks, sensor networks, crowdsourcing, real-time mobile systems, open-data, new analytics software).

1.3 The Characteristics of Big Data

Organizations gather more data than ever, faster and at lower cost. This data can be human-generated via use of smartphones, social networks or web surfing for example, or can be machine-generated, using RFID tags, surveillance cameras, or satellite imagery. This data can be further divided in two categories: Big Data in motion and Big Data at rest (Olofson & Vesset, 2012). Big Data in motion represents rapidly streaming, high-volume data that must be harnessed and synthesized as it arrives. Big Data in motion requires technology that can support real-time data analytics. Big Data at rest involves the ability to collect, process, and analyze the data and saving it in a state in which meaningful search, mining, discovery, query, and reporting may be conducted afterwards. For example, "a retailer analyzes a previous month's sales data and uses it to make strategic decisions about the present month's business activities. The action takes place after the data-creating event has occurred" (Internap, 2013).

Big Data technologies address bottlenecks faced by traditional business intelligence (BI) solutions designed for "normal" data gathered and structured with specific usages in mind. Often, uses for Big Data are determined after it is gathered. One could say BI data is "profiled and filtered by design" while Big Data is generally "filtered and profiled after the fact". Data profiling is the process of analyzing incoming data to detect problems, clean the data, and filter data that may have value in order to use them for Big Data analytics. Big Data analytics is the process of examining large amounts of data of a variety of types to uncover hidden patterns, unknown correlations and other useful information (TechTarget, 2012). These processes are particularly powerful at employing unstructured Big Data on the move.

1.4 Examples of Big Data

Big Data touches almost every application domain and more and more initiatives are showcased in literature. Examples occur in Big Sciences (e.g., astronomy, genomics, physics) as well as in various domains in line with Canada's priorities (Government of Canada, 2013 and 2013b) such as natural resources, infrastructure, health, security and defense.

In astronomy, the Sloan Digital Sky Survey (SDSS) collected more data in its first weeks of operation than had been collected in the entire history of astronomy. The Large Synoptic Survey Telescope program will start in 2016 and will acquire in 5 days the same amount of data the SDSS did in 10 years (Mayer-Schönberger & Cukier, 2013). In physics, at CERN, the Large Hadron Collider stores 200 petabytes per year after filtering 99.999% of the data. Without filtering, it would produce 500 exabytes per day, i.e., 200 times higher than all the other sources combined in the world (Wikipedia, 2014).

The "Waterfront Toronto" project includes a platform that integrates multiple data sources (e.g., sensors) and creates real-time visualization of information including traffic congestion reports, public transit information, and weather. Additional Big Data analytics capabilities within the platform will provide deeper insight around wellness, transportation, energy management, water conservation and sustainability efforts, as well as public safety and security across the community (IBM 2013).

In health, IBM (2012) reports that Premier, the largest United States healthcare alliance (2,700 hospitals, 100,000 non-acute care facilities and 400,000 physicians) provides physicians unprecedented access to best practices and can match patient care protocols with clinical outcomes to improve patient care. Since 2008, participating hospitals have saved an estimated 92,000 lives while reducing healthcare spending by 9 billion dollars. In another project, the Illinois Department of Healthcare and Family Services used Big Data analytics to detect overpayments fraud at a high accuracy level. The Hospital for Sick Children of Toronto also integrates large volumes of physiological readings from monitoring equipment that allow clinicians to detect life-threatening infections days sooner than the previous techniques that were used (Blount, et al, 2010).

1.5 The Players and Technologies of Big Data

The main players in the world of Big Data are large hardware, software and cloud computing providers (e.g., IBM, Oracle, Google, Amazon), social networks (e.g., Facebook, Twitter), research laboratories (e.g., NASA) and service companies (e.g., retail, banking, insurance). While the latter are highly regulated with regards to privacy, the former are less regulated and many offer Big Data analytics services.

The technology underpinning Big Data uses web-based distributed and scalable applications and frameworks that rely heavily on the Hadoop 2.0 ecosystem from the Apache Software Foundation. This technological ecosystem provides the required solutions to handle the large and high-velocity workloads of Big Data. Several products are developed to improve certain aspects of this ecosystem, for example the highly popular YARN (The Apache Software Foundation, 2013) and REEF (Retainable Evaluator Execution Framework, from Microsoft) solutions (Sears, 2013). However, not all Big Data technology is open source (Soares, 2013). Providers of proprietary solutions such as IBM, SAP, SAS, Oracle, Teradata and others play significant roles in the evolution of uses of Big Data. This global ecosystem delivers new possibilities at affordable costs while complementing existing BI structures that remain the major asset of organizations.

2. Big Data and Geomatics

With the many challenges facing Society at multiple scales, location has emerged as a key factor in decision-making (Rajabifard & Coleman, 2012). The next sections present how location information technologies contribute to the Big Data phenomenon and, inversely, how Big Data technology contributes to geomatics – the science and practice of activities related to location information capture and processing; location information analysis and presentation; integrated information products and services; and provision of location-based solutions . Then, we discuss the challenges and opportunities for the geomatics community with regards to Big Data.

2.1 The Contribution of Geomatics to Big Data

Big Data is enriched location information captured from GPS-enabled devices such as smartphones, cameras and on-board navigation systems in vehicles, as well as georeferenced sensors networks, for example traffic measurement devices, location-aware social networks, weather stations, cell phone towers, security surveillance cameras, RFID tagged supply chains, and soon Google Glass and other wearable devices such as smart watches. Geomatics contributes to the success of Big Data in three of the following ways:

2.1.1 More Powerful Analytics

Geospatial, or location-based data includes the position in a given space as well as the shape, orientation and size of phenomena. This data allows for querying about the distance, direction, height difference, shortest path and other properties. It also allows for analyzing spatial relationships (e.g., adjacency, connectivity, inclusion, proximity, exclusion, overlay, etc.) and spatial distribution (e.g., concentrated, scattered, grouped, regular, etc.). When time is added to space, we can also study the movement of phenomena, including the merging and splitting of located objects, the duration and number of presences in a place, the frequency and places where two phenomena overlay, the occurrence of spatio-temporal clusters and the spatial trends for given periods. In other words, adding spatial and temporal references leads to the creation of new analytical possibilities and the discovery of new facts, which can contribute significantly to the information derived from Big Data analytics.

2.1.2 Integration of Unlinked Big Data

Location reference can serve as an integrator, a comparator, an aggregator. Location-based data help to provide integrated local, regional, national and global insights. A large number of Big Data sources include a spatial reference (at least a location in 2-dimensional coordinates, e.g., latitude-longitude). Using the spatial reference allows for the integration of independent Big Data sets having nothing in common besides location, even if they use very different spatial referencing methods. Specialists in geomatics can transform them and bring them in a unique spatial reference system, opening the way to meaningful integration of independent datasets. In addition, it is possible to use the spatial reference to enrich Big Data sets with traditional geospatial data sources such as topography, road networks, census tracks, administrative boundaries and other potentially useful data. Since spatial and temporal references remain the only potential commonalities between independent systems, independent Big Data that would otherwise be unlinkable can be integrated *a posteriori*. According to (Agrawal, et al., 2012), the value of all data explodes when it can be linked with other data.

2.1.3 Enriched Data Visualization

Maps – the most common way to present location information - reveal insights that tables and statistical charts do not (e.g., clusters across administrative units). Spatial reference provides more efficient ways to visualize and analyze Big Data, typically using various types of 2D maps, 3D perspective views and profiles, or immersive animated 3D navigation in real-life augmented-reality environments. When integrated with temporal references, a series of maps and videos can support the visualization of the evolution of phenomena. These are natural aids to the analytical process. In the context of data exploration, maps do more than just make the data visible, they are active instruments to support the end-user's thinking process (MacEachren & Kraak, 2001). Consequently, Big Data projects using geovisualization techniques provide more insight to decision-makers and facilitate the analysis of geographically distributed phenomena.

2.1.4 Examples of the Contribution of Geomatics Sciences to Big Data

Web-based businesses like Facebook, Google and Yahoo! already use Big Data analytics to link user location with activities, to predict locations for a particular activity, and to predict activities at a given location. Retailers that use location analytics for logistics and tracking of their own inventory include Walmart, eBay, Amazon, Costco, Home Depot, and Sears, among others. Within governments, transportation, public works, planning, and economic development operations at all levels collect large amounts of location-related data through online services, taxpayer/customer reports, and vehicle-based sensors. Location-based service providers such as TomTom use Big Data analytics to provide more reliable real-time traffic advisory information to users of their premium subscription services. Similarly, power utilities use real-time customer demand information from the georeferenced sensors inside "smart meters" to balance their networks. There are many more examples of the integration of location information and Big Data.

Some of these applications use location data as accessory data that only requires simple spatial processing. These Location-Aware Big Data highly benefit from location data but do not require a high level of geomatics expertise. In contrast, other usages of Big Data consider geospatial information as central, and they go beyond point coordinates by providing information about shapes, textures, sizes, directions, connectedness, inclusion, motions, expansions, contractions, merging, divisions, etc. They first of all look for geometric and geospatial characteristics of phenomena and the geographic scales monitored can range from a single room to an entire city or ecosystem.

An example of Big Data relying on earth observation information include massive production of high-resolution satellite imagery such as the NASA EOSDIS program, which yields 5 terabytes of Earth data per day (Baumann, 2013). Another example is the European EarthServer project, which provides access to very large volumes of varied multisource, multi-dimensional, spatio-temporal data aims at providing “Big Earth Data Analytics” (Percivall, 2013). This Big Data project also supports a wide variety of data sources, from smartphones to immersive virtual reality, Triangulated Irregular Networks (TINs), vector sources, raster sources, image time series, point clouds, trajectories, meshes, solids, and more.

In the field of forestry, Global Forest Watch (GFW) fights against deforestation by providing timely information on the state of the world’s forests (World Resource Institute, 2014). GFW combines near-real-time satellite data, forest management and company concession maps, protected areas maps, mobile technology, crowd-sourced data, and on-the-ground networks and seeks to promote transparency in forests around the world.

Most Big Earth Data are Big Data at rest, but one can expect more real-time high-velocity projects in the near future, such as the Tango project from Google, which aims at giving mobile devices a human-scale understanding of space and motion, thanks to sensors allowing smartphones to make over a quarter million 3D measurements every second, updating their position and orientation, and combining data into a real-time 3D model of the world around users for various innovative usages (Google, 2014).

Other examples include the Amazon drone delivery system project (Amazon, 2013), the MIT SkyCall project (MIT Senseable City Lab, 2013), ocean and forest monitoring, real-time mobile mapping and augmented reality, LIDAR data sets, large 3D point clouds and high-density georeferenced sensor networks, to name a few. Such projects can be qualified as Big Earth Data (Open Geospatial Consortium, 2013).

2.2 The Contribution of Big Data to Geomatics

The Big Data phenomenon contributes to Geomatics in two ways: as a facilitator, and as a source of innovation.

The Big Data technological ecosystem facilitates the storing and processing of geospatial data, in particular with the use of “big processing” through cloud computing and analytics. Cloud initiatives have provided the geomatics industry with access to powerful, scalable storage and processing services hosted at remote locations. From this point of view, it simplifies the work already done in geomatics.

On the other hand, Big Data (along with Business Intelligence (BI)) concepts are impacting the geomatics community by leading to new geospatial-specific solutions, new bodies of knowledge, new scientific communities, new specialized conferences, and new working groups in standardization bodies (e.g., OGC). In recent years, we have seen communities emerging with different but complementary bodies of knowledge and tools. These communities use different names depending upon their technological roots and the problems they focus on. For example, the *Location Intelligence* community focuses on combining analytics capabilities with the transactional approach of Geographical Information Systems (GIS) to provide new insights in spatially-referenced business data. On the other hand, the *Geospatial Business Intelligence* community (or *GeoBI*) focuses on adapting the BI analytical data structures and operators to provide decision-makers with interactive multi-scale spatio-temporal data exploration, especially with Spatial OLAP (SOLAP) and spatial dashboards technology. We sometimes see *Location Analytics* for the same objective. Finally, the *Geovisualization* community focuses on users' real-time interactions with visualization tools for large volumes of static as well as dynamic geospatial data. The three communities' innovative tools and methods have been influenced by the rise of the Big Data phenomenon. In spite of a high level of overlap and the common influence of Big Data (and BI), no consensus has been reached yet on a unifying term for these communities.

Finally, the Big Data technological ecosystem has contributed to the emergence of the concept of "Spatially-Enabled Society", a concept where location, place and other geospatial data are commonly available to governments, communities and citizens (Rajabifard & Coleman, 2012).

2.3 The Challenges of Big Data in Geomatics

This section addresses the five challenges of Big Data in geomatics: location privacy, embracing the new paradigm in geomatics, workforce geomatics skills, spatial interoperability, and spatial data processing technology.

2.3.1 Location Privacy

Location privacy is the primary challenge for Big Data. The use of spatial referencing in Big Data raises new issues rarely discussed in the traditional Big Data literature. Positions, trajectories, speeds, breaks, cycles, and other movement data provide hints that are only beginning to raise additional concerns. More than any other type of Big Data, geospatial data has the potential to break traditional anonymized data and help to reveal the identity of individuals. This inherent capability has important implications (de Montjoye, Hidalgo, Verleysen, & Blondel, 2013). Because of their unique data integration and analytical capabilities, special care must be taken. For example, the usual practice of decoupling Big Data from identifiers (e.g., user's name) is not sufficient since it has been demonstrated that anonymized location data can be linked to individuals with 95% success if correlated with other data because our movement patterns are predictable (de Montjoye, Hidalgo, Verleysen, & Blondel, 2013).

The providers of location-based services can gain an intimate overview of one's habits, even if this information was never given to them, by building profiles from repeated movements to the same places, repeated inactivity, and periodical spatial clusters (e.g., where an individual works and goes shopping, his favorite hobbies, his religious habits, his participation in political activities, his travel preferences, his fast-food and sleeping place habits, his stops at drugstores). Big Geospatial Data can also help to infer networks of friends from spatio-temporal analytics (e.g., same place at the same time). Similarly, RFID is used in Spatially-aware Big Data projects to track people's movements in stores. Today's geospatial technology can use simple three-axis accelerometers (as used in smartphones) to discover the gender, size, weight and gait of individuals and successfully identify them (Meyer, 2013). In the Wall Street Journal, Thurm & Kane (2010) revealed that half of the top 100 apps for smartphones were disclosing users' locations to third parties without their consent. Big Geospatial Data also raises concerns about the privacy of the commercial content of containers, vehicles and cargos as their trajectory can be analyzed between warehouses or ports and deductions made. In fact, the strengths of geospatial data identified in section 2.1 (i.e., more powerful analytics, integration of unlinked Big Data, enriched data visualization) bring their own challenges to privacy. Geospatial data is much more powerful than non-geospatial data; consequently, Big Geospatial Data projects must be more careful!

Accordingly, new regulations are proposed, such as the US Location Privacy Protection Act (proposed in 2011), and modifications are proposed to existing laws and acts (Soares, 2013; Scassa, 2009). The point is to balance privacy issues with the potential gains for society. In fact, people want better data privacy but keep giving up their location data without much consideration. "We are not going to stop all this data collection, so we need to develop workable guidelines for protecting people. Those developing data-centric products also have to start thinking responsibly" (Meyer, 2013).

2.3.2 Embracing the New Paradigm in Geomatics

The geomatics community must accelerate its move beyond traditional GIS and mapping. The move is just beginning towards Geospatial Business Intelligence (GeoBI) in spite of the availability of commercial solutions since 2005 and scientific publications since the end of the 1990s. The GeoBI, Location Intelligence and Geovisualization communities are small in comparison with the global geomatics community. There is an opportunity for the geomatics community to embrace almost simultaneously the GeoBI and Big Data paradigms since they are closely related.

2.3.3 Workforce Geomatics Skills

Developers and users of Spatially-aware Big Data projects and of Big Earth Data projects need a sound knowledge of the spatial nature of the data because: using non-geospatial methods for geospatial data analysis may lead to erroneous results (e.g., distances and clusters not taking geographic obstacles into consideration); ignoring spatial reference systems may lead to erroneous positions; using data produced by cartographic generalization processes may lead to strange results when integrated with more precise location data (e.g., GPS); using 2D map measurements of 3D phenomenon (e.g., roads) can create analysis errors; etc. Big GeoData Scientists require skills in geomatics sciences, mathematics, geostatistics, programming, geospatial IT ecosystems, data governance, and geospatial data quality. Furthermore, they need a different mindset (e.g., creativity, empathy, sense-making) as they have to combine skills and talents with disjoint areas of expertise (e.g., with sociology) (Wachowicz, 2013). A geomatics engineer can be this kind of person but deeper analytical knowledge is needed (Wachowicz, 2013). Although some needs are being filled with the University of New Brunswick Cisco Geomatics Research Chair in Big Data Analytics, the recent NSERC Industrial Research Chair in Geospatial Business Intelligence at Laval University, and the private Research Chair in GeoBusiness at Sherbrooke University, insufficient numbers of qualified people are being produced. No technology is efficient without highly-qualified personnel.

2.3.4 Spatial Interoperability between Geospatial Big Data Systems

Variety in reference systems can be handled by spatial data processing software and interoperability standards. However, some differences cannot be controlled totally when interoperating with multiple systems: diversity of measuring instruments, imprecision of measurement methods and tools, data acquisition specifications evolving over the years, independent data update policies, conflicting priorities over data quality, conflicting legal restrictions over the use of data, and so on. To efficiently extract meaningful insights from large, varied and high-velocity Spatially-aware Big Data and Big Earth Data, we must aim at spatial interoperability between such systems and deal with these issues. Standards are evolving in this direction to facilitate the discoverability and machine ingestion of geospatial Big Data via interoperable services (Open Geospatial Consortium, 2013). Enriched metadata, profiling and spatial analytics are also needed.

2.3.5 Spatial Data Processing Technology

Due to the explosive growth and diversity of geospatial data, traditional geospatial analysis tools (e.g., GIS) are often not adequate for the interactive analysis and geovisualization of the huge volumes of varied spatio-temporal high-velocity data. Improved technologies are needed to fully benefit from Spatially-aware Big Data and from Big Earth Data. Geographic Knowledge Discovery (GKD) offers important directions in the development of a new generation of geospatial analysis tools in these data-rich environments (Han & Miller, 2009). The same holds for Spatial OLAP, Spatial Dashboards, Geovisualization and Location Intelligence technologies. These technologies are maturing, but several challenges remain to meet all 3 Vs of Big Data. Spatiotemporal objects and relationships tend to be more complex than the objects and relationships in non-geographic databases, thus more CPU-intensive. Geospatial data profiling and aggregating technology needs to improve. In addition, spatio-temporal data indexes must meet the requirements of Big Data on the move. Consequently, developing scalable tools for extracting spatiotemporal rules from collections of diverse geographic data is a major GKD challenge (Han & Miller, 2009). To harness the full power of Spatially-aware Big Data and of Big Earth Data, we need to meet the challenge of developing new technologies.

2.4 Opportunities

With so many sensors on humans and physical objects (cf. Internet of Things) and so many connected phones, tablets, PCs and recently wristwatches producing data, “we are instrumenting the universe” (Hoskins, in (Bertolucci, 2013)). Millions of dumb objects are becoming smart devices, millions of people are becoming data generators, cities are becoming smart... a data-centric universe is on the rise. Hoskins explains that “This process of digitizing the world’s physical objects may prove to be the defining element of the age of data. All the objects in the world are going to become alive and Internet-connected in a way that they were not before”. He calls the new era the “age of data ubiquity”. According to him, hardware had its 20- to 30-year run, then software had its 20- to 30-year run, and we are now at the beginning of the data era. Such a view of the new web is shared by many visionaries. We believe that opportunities for the geomatics community exist on both the supply and the demand sides.

2.4.1 On the Supply Side

Experts in geomatics know how powerful geospatial data are with regards to integration, analytics and visualization. Engineers in geomatics have the knowledge to make the links between spatial and temporal reference systems and measurement methods, to consider precision issues, and to calculate the impacts of spatio-temporal distortions on Big Data analytical results. In fact, each challenge of section 2.3 provides opportunities: developing solutions to protect location privacy, training the users and the workforce with more analytics and geomatics skills respectively, facilitating the interoperability between Big Geospatial Data ecosystems, and developing new technologies. Additional opportunities exist, such as offering new data cleaning services, new multisource data integration services, new map-based decision-support analytics technologies, new real-time data integration/aggregation services, new expertise on Big Geospatial Data quality, etc. Complex governance issues will also require new services, not only for the more global Big Earth Data projects that cross borders and which require Big Geospatial Data centers, but also for Spatially-aware Big Data. With regard to new outsourcing services, cloud computing services for Big Geospatial Data must likely be offered by Canadian companies to respect national security rules and laws. In general, downstream services may offer mass processing for certain categories of users while more sophisticated users could run ad-hoc processing on their platforms (European Space Agency, 2013).

These opportunities require investments in research and development. The following list is only a subset of R&D opportunities which could open new markets and help create start-up companies:

- Develop Big Geospatial Data pre-partitioning techniques to facilitate in-memory processing.
- Develop spatially-partitioned, pre-aggregating, pre-integrating optimization methods to facilitate Big Geospatial Data analytics.

- Develop rapid spatial synthesis methods for higher level analytics without requiring perfect integration of lower-level spatial data (i.e., to satisfy the velocity criteria, immediately provide aggregated value estimates at the expense of perfect data that takes longer to obtain).
- Develop spatial synthesis methods which are "statistics/semantics-fidelity-oriented" as opposed to the traditional "geometry-fidelity/map-readability-centered" map generalization methods (Bédard, Rivest, & Proulx, 2007).
- Discover new interactive visualization methods for high-velocity Big Geospatial Data.
- Discover new spatio-temporal analytics methods for Big Geospatial Data.
- Extend spatio-temporal query languages for synthesis/summarization operations.
- Develop enriched metadata engines and services for Big Geospatial Data projects.
- Develop spatial extensions to traditional Big Data technologies to allow for storing, transforming, visualizing and analyzing basic geospatial data. Examples include:
    - richer nominal locators (cf. using place names with gazetteers);
    - explicit spatial relationships (e.g., city X within county Y within province Z; county A adjacent to county B; municipality X crossed by highway 20);
    - historical evolution (e.g., county X divided in Y and Z; municipality A changed from county Y to Z in 2008);
    - cross-border semantics harmonization (e.g., definitions of hydrographic features among provinces' topographic datasets in Canada (see Brodeur, 2003 for examples));
    - expanded quality metadata and build Big Geospatial metadata engine;
    - automatic error detection and correction; and
    - Big Geospatial Data interoperability quality assessment (Sboui, Bédard, Brodeur, & Badard, 2009).

Some geospatial products already work with the Hadoop ecosystem such as SpatialHadoop from the University of Minnesota and ESRI's GIS tools for Hadoop. To remain up-to-date with such topics and upcoming standards, one can follow the OGC Big Data DWG's open forum, which encourages collaboration and ensures liaison to other Big Data working groups inside and outside OGC (Baumann, 2013).

2.4.2 On the Demand Side

On the demand side, certain domains are adopting Big Data at a faster rate. By combining the 3 Vs of Big Data with the potential of certain markets with regards to geospatial data, we expect the following domains to be the early adopters:

- Geomarketing;
- Security;
- ITS (Intelligent Transportation Systems);
- Emergency disaster management;
- E911;
- Public health (epidemiology);
- Infrastructure monitoring and maintenance (e.g., pipelines, airports, hospitals);

- Supply chain management of commodities (e.g., food, water, oil, minerals);
- Smart City projects;
- Environmental monitoring; and
- Transparency of government and industry activities (i.e., public scrutiny in part due to increased availability of geospatial data).

All these market segments are interested in employing solutions that will help them better analyze and make predictions from the spatially-referenced information being received.

Generally speaking, while industry will succeed by offering new products and services, government will succeed by establishing new policies and governance (e.g., privacy, security, consumer protection) in addition to offering new services. Regarding post-secondary education, institutions will succeed by offering new curricula and research chairs. Considering Canada's recognized strength in higher education in geomatics engineering and geospatial information technologies, the Big Geospatial Data challenge is an opportunity in the international market.

3. Conclusions

Society is entering a data-centric era at a rapid pace. The number of devices connected to the Internet exceeded the number of people on Earth five years ago and is rapidly increasing. Very soon, we will collect more data in a single day than we have collected in our entire history. An important part of the Big Data discourse is similar to the discourse of BI. However, Big Data is not BI with bigger data. Big Data often comes from outside, typically from the cloud, and the main differences are in velocity and variety. These require new capabilities that are being addressed by Big Data core technologies as well as Big Data-enabled technologies.

Location and spatial referencing are ubiquitous components of Big Data projects, either as complementary information (cf. Spatially-aware Big Data) or as primary information (cf. Big Earth Data). Geospatial information technologies such as GPS, satellite imagery and web mapping contribute significantly to such systems. Although the early leaders are the large hardware, software, browser, social network, and retail companies, most organizations will want to benefit from the analytical power of Big Data. They will do it internally or with the help of external services. Considering that most projects use geospatial data, this represents a major opportunity for the geomatics industry.

Central to successful Big Geospatial Data projects are profiling, analytics and governance. Although these already exist for normal geospatial data, Big Geospatial Data projects add new components to the technological ecosystems already in place. Properly integrating these new components into existing geospatial ecosystems is a key ingredient to success.

Of particular importance is the explicit inclusion of location privacy issues in a project governance framework. Privacy is already a concern in our digital age and several solutions have been put in place within systems and in legislation. Nevertheless, while the very nature of privacy considerations is changing with the use of social networks and web services, Big Data geospatial capabilities raise privacy concerns to a new level. “With Big Data comes big responsibility” (Ramirez, 2013).

The difficult part of Big Data is not technological, it is figuring out what to do with the results. Big Data relies on the power of large numbers of facts rather than relying on the traditional causal or deductive relationships. With Big Data analytics, one must believe numbers even when the underlying reasons for these numbers remain unknown. Since it is too easy to mistake correlation for causation and to find misleading patterns in the data, knowledge in Big Data analytics, geospatial analysis, spatial data quality, and the field of application is required. The cultural changes are enormous and this paradigm shift represents a challenge for system designers, decision makers and data analysts. Fulfilling the demand for a workforce with the necessary skills is also a major challenge for post-secondary geomatics institutions.

So far, there has been a slow penetration of Location Intelligence, Geospatial Business Intelligence and Geovisualization technologies into the geomatics community. It is still limited to early adopters. Similarly, it appears that the major bottleneck for the geomatics community to embrace the Big Data market will be to accept going beyond the traditional GIS-DBMS-driven “data transaction culture” and going towards the analytical paradigm.

Technological trends emerge within a given context. Big Data emerged in an era involving the proliferation of affordable geospatial technologies, the Semantic Web, the Internet of Things, cloud computing, Smart Cities, data governance, and BI and analytics. According to Gartner (2013), every new technology begins with a period of excitement where hopes are enormous and technological challenges seem solvable. This period of hope is typically followed by a period of disillusion where early results deliver less than hoped for because of the complex coexistence with infrastructures in place, a high learning curve, deceptive results, etc. Then, expectations become more realistic, human resources more knowledgeable, and integration within existing technological ecosystems easier. This is when a technology becomes mainstream, the tools and resources more mature, and returns on investments (ROI) the most obvious. Although the hype about Big Data may be at its peak (LeHong & Fenn, 2013), Big Data is here to stay and geospatial information technologies will play a central role into its success and usefulness. Big Data is not only an exceptional marketing term, it brings a new attitude about embedding powerful analytics into organizations and getting additional value from what the spatially-referenced data can tell us. Accordingly, Big Geospatial Data offers plenty of new opportunities.

The winners in the global economy will be those countries with the right analytical brains and tools to develop a better understanding of human-related activities, environmental changes, health-related phenomena, market evolutions, etc. They will be in a better position to provide their citizens with increased security, timely healthcare, environment-friendly transportation, well-maintained infrastructures, faster-adjusting economies, and a better quality of life.

Appendix A: List of Acronyms

The following table presents the meaning of acronyms used in this document.

Table 1: List of Acronyms

Acronym/abbreviation	Meaning
2D	Two-Dimensional
3D	Three-Dimensional
BA	Business Analytics
BDMS	Big Data Management Systems
BI	Business Intelligence
CDA	Confirmatory Data Analysis
CEP	Complex Event Processing
CPU	Central Processing Unit
DBMS	Database Management System
DWG	Data Working Group
EDA	Exploratory Data Analysis
EOSDIS	Earth Observing System Data and Information System
FTC	Federal Trade Commission
GeoBI	Geospatial Business Intelligence
GFS	Google File System
GFW	Global Forest Watch
GIS	Geographic Information System
GKD	Geographic Knowledge Discovery
GPS	Global Positioning System
HDFS	Hadoop Distributed File System
HEC	Hautes Études Commerciales
ICT	Information and Communication Technologies
ISO	International Organization for Standardization
IT	Information Technology
ITS	Intelligent Transportation System
LIDAR	Light Detection and Ranging
MIT	Massachusetts Institute of Technology

Acronym/abbreviation	Meaning
NASA	National Aeronautics and Space Administration
NCCS	NASA Center for Climate Simulation
NSERC	Natural Sciences and Engineering Research Council of Canada
OGC	Open Geospatial Consortium
OLAP	On-Line Analytical Processing
OSM	OpenStreetMap
R&D	Research and Development
RAM	Random-Access Memory
RDA	Resource Description and Access
RDF	Resource Description Framework
REEF	Retainable Evaluator Execution Framework
RFID	Radio-Frequency Identification
ROI	Return on Investment
SDSS	Sloan Digital Sky Survey
SOLAP	Spatial On-Line Analytical Processing
SQL	Structured Query Language
TIN	Triangulated Irregular Network
URI	Uniform Resource Identifier
US	United States
W3C	World Wide Web Consortium
XML	Extensible Markup Language

Appendix B: Big Data Definitions

This appendix summarizes Big Data characteristics, key operations with Big Data, and governance considerations.

Key Data Characteristics

The key characteristics, or the 3 “Vs”, of Big Data: “Volume”, “Velocity”, and “Variety” were introduced in the previous section. Each is discussed here in more detail.

Volume

Volume refers to the amount of data to be handled. It can be measured by the quantity of transactions, events, or amount of history, and further amplified by their attributes, dimensions, or predictive variables (Minelli, Chambers, & Dhiraj, 2013). In 2020, we will reach 35 zettabytes of information (1ZB = 1 billion terrabytes) (Eaton, Deutsch, Deroos, Lapis, & Zikopoulos, 2012). Current statistics include:

- Twitter users are sending 400 million tweets per day (estimated to be more than 7 terabytes), which represents 146 billion tweets per year and growing (Karel, 2013).
- Facebook gets more than 10 million photos every hour and Facebook members click a “like” button or leave a comment 3 billion times per day (Mayer-Schönberger & Cukier, 2013).
- In 2012, Google reported over 5 billion searches per day (Karel, 2013), processed more than 24 petabytes of data per day and had 800 million YouTube users per month (Mayer-Schönberger & Cukier, 2013).
- Walmart collects more than 2.5 petabytes of data every hour from its customer transactions (McAfee & Brynjolfsson, 2012).
- The National Aeronautics and Space Agency (NASA) Center for Climate Simulation (NCCS) stores 32 petabytes of climate data in its system composed of 35,000 processing cores calculating more than 400 trillion floating-point operations per second (Webster, 2012).
- OpenStreetMap (OSM) has relied on thousands of volunteers since 2004 to build a world-wide database of road features. The large volume of data gathered in the 10 years since the creation of OSM corresponds to the amount of volume gathered by Apple Computer in a single day (Apple Computer, 2013). This illustrates the difference between today’s normal data (OSM) and today’s Big Data. They are clearly separated by orders of magnitude in the volume of data as well as in the timeliness of the data.

In 2000, only 25% of the world information was digital. In 2007 it was 93% and it is now at 98%; furthermore, 90% of the world data is less than two years old (Wessler, 2013).

Variety

Variety refers to the number of different sources from which organizations can obtain/collect data. Many of these sources are automated (e.g., heat sensors) and are not designed to be user-friendly, so the data they collect is not necessarily in a format ready for analysis. Some sources provide data that is generated with no specific question in mind (e.g., tweets) or are a by-product of another activity (e.g., products' client segmentation from web surfing logs). Contrary to traditional sources that were structured by design, many of the new data sources are semi-structured (e.g., web surfing logs, XML documents) or unstructured (e.g., emails, clickstream data, videos), capture everything (that may be of use or not), and may be messy and noisy. Approximately 80 percent of the world's data is unstructured, growing at 15 times the rate of structured data (Eaton, Deutsch, Deroos, Lapis, & Zikopoulos, 2012).

Velocity

Velocity refers to both the data collection rate and analysis timeframe (Agrawal, et al., 2012). Velocity varies from batch integration and loading of data at predetermined intervals to real-time streaming of data (Olofson & Vesset, 2012). High velocity has an impact on accessibility (i.e., providing users with proper analytics information when, where and how the users want it). For example, Google's analysis of Internet users' queries about the H1N1 virus allowed for publishing detailed information about the worldwide spread of the virus two weeks before the traditional reporting method based on physicians' offices (Mayer-Schönberger & Cukier, 2013).

The Remaining 5 Vs

Additional characteristics are highlighted in the literature: "Value", "Validity", "Veracity", "Vulnerability" and "Visualization". **Value** refers to both the cost of technology and the value derived from the use of Big Data. Big Data in motion has value when captured and analyzed in real-time while Big Data at rest has value only when it is used. Some authors state that all data should be kept since it potentially contains relevant patterns while others mention that much of this data is of no value (i.e., noise), and it should be filtered before being stored. **Validity** refers to the technical correctness of the data (e.g., number vs text). **Veracity** refers to trustworthiness, provenance, lineage, and quality (Baumann, OGC Discussion, 2013). As we went from a low number of data sources of managed and known quality to millions of sources of unknown quality (e.g., blogs), issues of internal quality (e.g., errors, uncertainties) and external quality (e.g., usability, fitness-for-use) must be addressed with a fresh perspective, notably, with respect to legal issues regarding privacy, traceability of data, liability of providers and integrators, copyright, and public protection. **Vulnerability** refers to the level of protection of the data through its lifecycle. In a few application domains such as healthcare and finance, rules and regulations are in place to ensure the protection of data at rest as well as during execution of application logic or during transmission. **Visualization** refers to the graphical representation of data to communicate information clearly and effectively in spite of the high volume of data, its velocity and its variety. Numerous techniques exist (see Friendly, 2009; Few, 2012 and 2013).

Key Operations

Big Data Profiling

“Data profiling” is the initial phase of data integration projects. It helps to understand available data, to identify problems by analyzing content, structure, and relationships within and across datasets (e.g., number of missing fields and nulls, percentage of unique values, inconsistencies and redundancies). Data profiling tools for Big Data have a similar function to data profiling tools for normal data. In fact, data quality/cleaning/cleansing/scrubbing software used in BI and other projects is now being extended to support Big Data profiling. Nevertheless, data profiling challenges increase when dealing with unstructured and high-velocity Big Data. To compose with this higher degree of complexity, multi-phase approaches to Big Data profiling are proposed (Chastain & Loshin, 2013), which allow for determining the value of Big Data sources for particular purposes before spending resources to fully profile them.

Big Data Analytics

“Data analytics” is the process of examining data to uncover hidden patterns, unknown correlations, trends, weaknesses, new segmentations, forecast models, predictive behaviors and other useful patterns. It includes exploratory data analysis (EDA), where new features in the data are discovered, and confirmatory data analysis (CDA), where existing hypotheses are proven true or false. Data analytics provide insight into datasets that would simply be too costly, difficult and time-consuming to analyze otherwise.

Big Data Analytics focuses on the extraction of useful insights and predictive analysis in real time, from the rapidly increasing volumes of new heterogeneous data. “It is logical to say Business Intelligence (BI), Business Analytics, and Big Data are progressive steps up the information maturity scale, and they are very closely related” (Zhu, 2013). Gartner predicts that by 2015, over 50% of Big Data solutions will use data in motion (i.e., data streams) from instrumented devices, applications, events or individuals (Laney, 2013).

Big Data Governance

Data governance establishes rules to enforce the proper use of data, privacy, quality control, security, etc. Without proper data governance for the whole lifecycle of Big Data, the integration of this data is very difficult to do correctly and has a high potential to pose a number of problems (Soares, 2013). Effective data governance can enhance the quality, availability and integrity of an organization’s data, thus improving the quality of analyses. It directly impacts the four factors any organization cares about most: increasing revenue, lowering costs, reducing risks, and increasing data confidence, or trust.

Appendix C: Big Data Technologies

This appendix summarizes key IT technologies that have contributed to the emergence of the Big Data phenomenon.

Semantic Web/Linked Data

The “Semantic Web” is an extension of the World Wide Web that provides a common framework allowing data to be shared and reused across applications, enterprises, and community boundaries. It includes common formats to integrate data from diverse sources, and a language, called Resource Description Framework (RDF), to record how the data relates to real world objects and to other data. RDF is a world-wide lingua franca connecting web syntax with formal definitions/ontology (i.e., real-life meanings or semantics). It allows users or machines to search for data in one database, and then move through a set of databases that are connected by their content and semantics to find other, related data. This collection of interrelated datasets on the web can be referred to as “Linked Data” (World Wide Web Consortium (W3C), 2013).

Internet of Things

The Internet of Things refers to uniquely identifiable objects and their virtual representations in an Internet-like structure (Wikipedia, 2013). It is a connection of nodes (i.e., data sources) through the web, where every node has its own Uniform Resource Identifier (URI). It is estimated that there will be over 50 billion such resources by 2020 (Dasgupta, 2013). Examples include utilities with smart meters, and healthcare with the deployment of remote health monitoring devices. This creates an unprecedented continuous stream of data.

Cloud Computing

Cloud computing is the movement of applications, services, and data from local storage to a dispersed set of servers and datacenters (Berry, 2009). Workloads are allocated among a number of interconnected computers acting as a single machine. It is dynamically scalable, meaning that the system can be expanded as needed, and it acts as a web service. It is a true facilitator for Big Data projects. “Private clouds” are exclusive to a specific organization; “Community clouds” are restricted to departments or groups of users; “Public clouds” are exposed to everyone. “Hybrid clouds” are private clouds that may use supplementary resources from public clouds if needed.

Smart Cities

A city is an interconnected system of systems based on three elements: infrastructure, operations, and people (IBM Corporation, 2013). A city can be defined as “smart” when these elements continuously communicate and interact together. The resulting distributed network of fixed and mobile intelligent sensor nodes, including citizens, wirelessly provides a wide range of real-time information about the physical infrastructure, services, and interactions between people for more efficient management of the city. Data delivered in real-time to citizens and appropriate authorities helps to anticipate problems, to resolve them proactively, and to coordinate resources to operate effectively. Smart city, the Internet of Things, and Big Data concepts go together.

Business Intelligence and Analytics

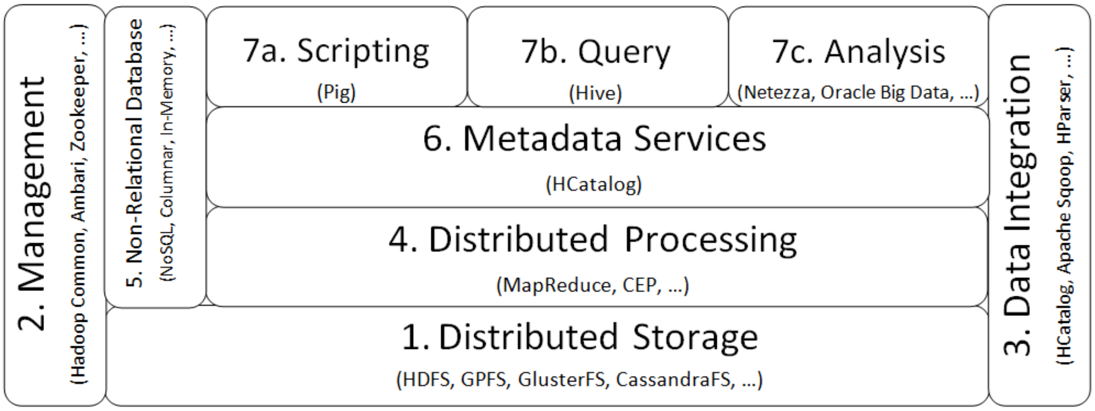
Business Intelligence (BI) is an umbrella term describing tools and methods to improve business decision-making by using fact-based support systems (Power, 2007). BI transforms large volumes of structured raw data into useful insights for strategic, tactical, and operational decision-making (Evelson, 2010). BI provides historical, current, and predictive views of business operations (Wikipedia, 2011). Over the last 20 years, BI has developed a strong data analytics culture, powerful data visualization solutions, as well as proven methods to integrate with organizations' structured database ecosystems. More recently, some authors started using the expression "Business Analytics" (BA) to focus on analytical and predictive functions. The BI market is hence evolving into two segments: the old, or traditional, BI market and the recent "data discovery" market (Hagerty, Sallam, & Richardson, 2012). Nowadays, BI is better tailored to meet the requirements of Big Data, such as support for large volumes of streaming data, real-time analysis and complex visualization. Accordingly, several practitioners see Big Data as the normal evolution of BI: they share a lot of commonalities in spite of new challenges that require new solutions. "Big Data is the next generation of data warehousing and business analytics." (Minelli, Chambers & Dhiraj, 2013). However, not everyone shares this point of view, especially for unstructured Big Data on the move.

Big Data Core Technologies

New technologies have emerged to address the exploding volumes of complex and high-velocity data. They have been developed specifically to capture the value of Big Data. One such technology is Hadoop, which is used to store and process large non-relational datasets via a large, scalable distributed model (The Apache Software Foundation, 2013). Hadoop has become a major trend, which was inspired by Google's work on its Google File System (GFS) and the MapReduce distributed data processing framework that provides highly parallel data processing and analysis capabilities. One of the key characteristics of Hadoop is the redundancy built into the environment. Data is redundantly stored in multiple places across the cluster, and the programming model is such that failures are expected and are resolved automatically by running portions of the programs on various servers in the cluster. Hadoop is an ecosystem of projects targeted at simplifying, managing, coordinating, and analyzing large sets of data (Eaton, Deutsch, Deroos, Lapis, & Zikopoulos, 2012). Several vendors offer Hadoop distributions.

The new Big Data-specific technologies touch many aspects of the data management framework. Figure 1 presents these layers and will help position the technologies discussed in this section.

Figure 1: The Building Blocks of a Data Management Framework (adapted from (Iron Systems, 2013))


2. Management
(Hadoop Common, Ambari, Zookeeper, ...)
5. Non-Relational Database
(NoSQL, Columnar, In-Memory, ...)
7a. Scripting
(Pig)
7b. Query
(Hive)
7c. Analysis
(Netezza, Oracle Big Data, ...)
6. Metadata Services
(HCatalog)
4. Distributed Processing
(MapReduce, CEP, ...)
1. Distributed Storage
(HDFS, GPFS, GlusterFS, CassandraFS, ...)
3. Data Integration
(HCatalog, Apache Sqoop, HParser, ...)

Distributed Storage (1) refers to how large volumes of information are stored and retrieved in the redundancy-based, failure-safe Big Data ecosystem. Common considerations include the storage of data for later access, the online vs. archival temporal extents, the archival vs. deletion temporal extents, etc. Management Services (2) refer to various utilities that support the operation of the data framework. Data Integration Services (3) are used to combine Big Data (as well as normal data) coming from various sources and in various formats. For example, retailers can combine, in real-time, selected clickstream data on the web with supply chain management RFID data, along with social network posts and "Likes", for the timely distribution of proper quantities of products to their stores. Distributed Processing (4) refers to how data is manipulated in the data management framework, and how meaningful events are processed. Non-Relational Databases (5) refer to the new types of database management systems (DBMS), sometimes called Big Data Management Systems (BDMS), that have been developed to address the various dimensions of Big Data (e.g., NoSQL, columnar DB, in-memory DB). Metadata Services (6) aim at documenting the characteristics, source and nature of data. Metadata is a fundamental, although often neglected, piece of Big Data. Scripting/Query/Analysis (7) is a family of tools (e.g., programming languages, query languages, analytical hardware-software combinations) allowing users and programmers to interact with Big Data.

Big Data core technologies explicitly address the synchronization between the systems that collect and host the data and the systems that perform analyses. When using traditional technologies, the process of preparing data for analytical purposes can be tedious and slow. The Big Data pipeline must be straightforward to meet the requirements of performance and automation. For Big Data in motion, analysis techniques that can process streaming data on the fly are needed, since it is not desirable to store the data first and analyze it afterward. In some cases, the ability to develop partial results in advance (e.g., by pre-aggregating samples of the data) so that only a small amount of incremental computation is necessary as new data arrives, can provide a good compromise.

Big Data Enabled Technologies

While new products were developed from scratch to meet the Big Data challenges, more traditional technologies have improved. Currently, these Big Data-enabled technologies address successfully 2 of the 3 key Vs (volume and velocity). They are still improving to efficiently support the 3rd V of Big Data (variety). For example, many data integration tools, DBMS and Scripting/Query/Analysis tools now provide bridges to Hadoop, some DBMSs better support semi-structured data, metadata tools have been extended to support Big Data, etc. From a user point of view, there are numerous advantages related to the use of these Big Data-enabled technologies: the basic technological concepts are already well understood; there already is legacy code available and tested (no need to redo it); the costs for the adapted technologies are often lower than the costs of acquiring new specific technologies; the security model is well under control; users are already proficient with these; and only minor additional training may be required (compared to the full training required on new technologies). It is expected that organizations will typically add Big Data technologies into a more global information ecosystem, with a selection of software tailored to cover all their traditional needs and their Big Data needs.

Appendix D: References

Agrawal, D., Bernstein, P., Bertino, E., Davidson, S., Dayal, U., Franklin, M., et al. (2012). *Challenges and Opportunities with Big Data.* Computing Community Consortium. Computing Community Consortium.

Amazon. (2013). *Amazon Prime Air.* Retrieved from Amazon: http://www.amazon.com/b?node=8037720011

Apple Computer. (2013). *Private meeting.* Cuppertino, CA.

Barnes, R. (2013, June 6). Big data issues? Try coping with the Large Hadron Collider. *Marketing*.

Baumann, P. (2013). *Big Data DWG Charter.* Open Geospatial Consortium.

Baumann, P. (2013). OGC Discussion.

Bédard, Y., Rivest, S., & Proulx, M.-J. (2007). Spatial On-Line Analytical Processing (SOLAP): Concepts, Architectures and Solutions from a Geomatics Engineering Perspective. In R. Wrembel, & C. Koncilia (Eds.), *Data Warehouses and OLAP: Concepts, Architectures and Solutions* (pp. 298-319). London: IRM Press.

Berry, J. K. (2009, September). GIS and the Cloud Computing Conundrum. *GeoWorld*, pp.12-13

Bertolucci, J. (2013, January 4). *The Age Of 'Data Ubiquity': Sensors Spread.* Retrieved from InformationWeek: http://www.informationweek.com/big-data/big-data-analytics/the-age-of-data-ubiquity-sensors-spread/d/d-id/1109327?

Blount, M., Ebling, M. R., Eklund, M. J., James, A. G., McGregor, C., Percival, N., et al. (2010, March/April). Real-Time Analysis for Intensive Care - Development and Deployment of the Artemis Analytic System. *IEEE Engineering in Medicine and Biology* , pp. 110-118.

Brodeur, J. (2003). Interopérabilité des données géospatiales: élaboration du concept de proximité géosémantique. *Thèse de doctorat*.

Caragliu, A., Del Bo, C. F., & Nijkamp, P. (2009). *Smart Cities in Europe.* VU University Amsterdam, Faculty of Economics, Business Administration and Econometrics.

Chastain, S., & Loshin, D. (2013). *How to Use an Uncommon-Sense Approach.* SAS.

Dasgupta, A. (2013, April). Big data: The future is in analytics. *Geospatial World*.

de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013, March 25). Unique in the Crowd: The privacy bounds of human mobility. *Nature*.

Eaton, C., Deutsch, T., Deroos, D., Lapis, G., & Zikopoulos, P. (2012). *Understanding Big Data Analytics for Enterprise Class Hadoop and Streaming Data*. McGraw-Hill.

European Space Agency. (2013). *Big Data from Space Event Report*. Rome, Italy.

Evelson, B. (2010). *Want to know what Forrester's lead data analysts are thinking about BI and the data domain?* Retrieved from Forrester Blog: http://blogs.forrester.com/boris_evelson/10-04-29-want_know_what_forresters_lead_data_analysts_are_thinking_about_bi_and_data_domain

Fenn, J., & LeHong, H. (2011). *Gartner Hype Cycle for Emerging Technologies 2011*. Gartner.

Few, S. (2013). *Information Dashboard Design: Displaying Data for At-a-Glance Monitoring* (2nd ed.). Analytics Press.

Few, S. (2012). *Show Me the Numbers: Designing Tables and Graphs to Enlighten* (2nd ed.). Analytics Press.

Franks, B. (2012). *Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics.* Wiley.

Friendly, M. (2009). *Milestones in the history of thematic cartography, statistical graphics, and data visualization.*

Gartner. (2013). Methodologies: Hype Cycles.

Google (2014). The Project Tango. http://www.google.com/atap/projecttango/

Government of Canada. (2013b). *Report for the United Nations Economic and Social Council.* Country Report of Canada.

Government of Canada. (2013, October 16). Speech from the Throne. Canada.

Hagerty, J., Sallam, R. L., & Richardson, J. (2012). *Gartner Magic Quadrant for Business Intelligence Platforns.* Gartner.

Han, J., & Miller, J. H. (2009). Geographic Data Mining and Knowledge Discovery: An Overview. In J. Han, & H. J. Miller (Eds.), *Geographic Data Mining and Knowledge Discovery Second Edition* (p. 458). CRC Press.

Han, J., Halevy, A., Giles, L., Leskovec, J., Hearst, M., & Bennett, P. (2013, October 31). Channeling the Deluge: Research Challenges for Big Data and Information Systems. *ACM Conference on Information and Knowledge Management (CKIM 2013) - Panel Discussion* . San Francisco, California, USA.

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.

IBM Corporation. (2007). *The IBM Data Governance Council Maturity Model: Building a roadmap for effective data governance.* IBM Corporation.

IBM Corporation. (2012). *Premier: Helping healthcare providers deliver the best possible care to their patients.* IBM Corporation.

IBM Corporation. (2013). *Smarter Cities.* Retrieved from A Smarter Planet: http://www.ibm.com/smarterplanet/us/en/smarter_cities/overview/

IBM Corporation. (2013, September 18). *Waterfront Toronto Teams with IBM to Build a Smarter City.*: http://www.ibm.com/news/ca/en/2013/09/18/d784454e42662t01.html

Iron Systems. (2013). *Kick start Hadoop with the right platform.* Retrieved from Iron Systems: http://www.ironsystems.com/products/hadoop-platforms-overview

Karel, R. (2013, November 18). Big Data, So Mom Can Understand. *Perspectives The Informatica Blog*.

Laney, D. (2001). *3D Data Management: Controlling Data Volume, Velocity, and Variety.* Meta Group.

Laney, D. (2013, November 14). Big Data and Analytics Strategy Essentials.

Laney, D. (2012). *The Importance of 'Big Data': A definition.* Gartner.

LeHong, H., & Fenn, J. (2013). *Gartner Hype Cycle for Emerging Technologies.* Gartner.

MacEachren, A., & Kraak, M.-J. (2001). Research challenges in geo-visualization. *Cartography and Geographic Information Science , 28* (1), pp. 3-12.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A revolution that will transform how we live, work and think.* Houghton Mifflin Harcourt.

McAfee, A., & Brynjolfsson, E. (2012, October). Big Data: The Management Revolution. *Harvard Business Review* , pp. 61-68.

McBurney, V. (2012, May 31). The Origin and Growth of Big Data Buzz.

McKinsey Global Institute. (2011). *Big data: The next frontier for innovation, competition, and productivity.*

Meyer, D. (2013, March 25). *Why the collision of big data and privacy will require a new realpolitik*. Retrieved from Gigaom: http://gigaom.com/2013/03/25/why-the-collision-of-big-data-and-privacy-will-require-a-new-realpolitik/

Minelli, M., Chambers, M., & Dhiraj, A. (2013). *Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses*. Wiley.

MIT Senseable City Lab. (2013). *SkyCall*. Retrieved from MIT Senseable City Lab: http://senseable.mit.edu/skycall/

Nixon, N. (2013, June 20). *Data in motion vs. data at rest*. Retrieved from Internap: http://www.internap.com/2013/06/20/data-in-motion-vs-data-at-rest/

Olofson, C. W., & Vesset, D. (2012). *Big Data: Trends, Strategies, and SAP Technology*. IDC Information and Data.

Open Geospatial Consortium. (2013, October 22). *BigDataDwg Web*. Retrieved from OGC Public Wiki: http://external.opengeospatial.org/twiki_public/BigDataDwg/WebHome

Open Geospatial Consortium. (2013, October). OGC Public Wiki - Big Data Dwg.

Percivall, G. (2013, August 8). Big Processing of Geospatial Data. *OGC Update Blog*.

Power, D. (2007). *A Brief History of Decision Support Systems, version 4.0*. Retrieved from DSSResources.com: http://dssresources.com/history/dsshistory.html.

Priestley, T. (2013, December 16). *Just what is a Data Scientist anyway ?* Retrieved from LinkedIn: http://www.linkedin.com/today/post/article/20131216094029-2143418-just-what-is-a-data-scientist-anyway

Rajabifard, A., & Coleman, D. (2012). Towards Spatial Enablement and Beyond. In A. Rajabifard, & D. Coleman (Eds.), *Spatially Enabling Government, Industry and Citizens. Research and Development Perspectives* (pp. 9-22). GSDI Association Press.

Ramirez, E. (2013). *The Privacy Challenge of Big Data: A View from the Lifeguard's Chair*. Keynote Address by the US Federal Trade Commission Chairwoman, Technology Policy Institute Aspen Forum, US Federal Trade Commission, Aspen, Colorado.

Sboui, T., Bédard, Y., Brodeur, J., & Badard, T. (2009). Modeling the External Quality of Context to Fine-tune Context Reasoning in Geospatial Interoperability. *The 21st International Joint Conference on Artificial Intelligence*. Pasadena, California, USA.

Scassa, T. (2009). Information Privacy in Public Space: Location Data, Data Protection and the Reasonable Expectation of Privacy. *Canadian Journal of Law and Technology, 7* (2), pp.193-220

Sears, R. (2013, October 28-30). REEF - Retainable Evaluator Execution Framework. *Strata Conference + Hadoop World* . New York.

Soares, S. (2013). *Big Data Governance: An Emerging Imperative.* Mc Press.

Taylor, J. (2013, October 22). Predictive Analytics in the Cloud 2013 - Opportunities, Trends and the Impact of Big Data.

TechTarget. (2012, January 12). *Big Data Analytics*. Retrieved from Search Business Analytics: http://searchbusinessanalytics.techtarget.com/definition/big-data-analytics

TechTarget. (2011). *Essential Guide - Building an effective data governance framework.* TechTarget.

The Apache Software Foundation. (2013, October 7). *Apache Hadoop NextGen MapReduce (YARN)*. Retrieved from Hadoop: https://hadoop.apache.org/docs/current2/hadoop-yarn/hadoop-yarn-site/YARN.html

The Apache Software Foundation. (2013, February 2). *HCatalog Table Management.* Retrieved from The Apache Software Foundation: http://hive.apache.org/docs/hcat_r0.5.0/

The Apache Software Foundation. (2013, December 11). *Welcome to Apache Hadoop!* Retrieved from The Apache Software Foundation: http://hadoop.apache.org/

Thurm, S., Kane, Y.I. (2010, Dec. 17). Your Apps are Watching You. *The Wall Street Journal*.

Wachowicz, M. (2013, October 28). New Frontiers for Geomatics - Harnessing the Smart City Space of Tomorrow. *GIM International, 27* (10).

Webster, P. (2012). Supercomputing the Climate: NASA's Big Data Mission. *CSC World*.

Wehbe, B. (2013, June 6). *Big Data: Is It Just Another Big Hype.* Retrieved from Enterprise Systems Media.

Wessler, M. (2013). *Big Data Analytics for Dummies.* Hoboken, New Jersey: John Wiley & Sons

Wikipedia. (2014, January 16). *Big data.* Retrieved from Wikipedia: Agrawal, D., Bernstein, P., Bertino, E., Davidson, S., Dayal, U., Franklin, M., et al. (2012). Challenges and Opportunities with Big Data. Computing Community Consortium. Computing Community Consortium.

Wikipedia. (2011). *Business Intelligence.* Retrieved from Wikipedia: http://en.wikipedia.org/wiki/Business_intelligence

Wikipedia. (2013, December 12). *Internet of Things.* Retrieved from Wikipedia: http://en.wikipedia.org/wiki/Internet_of_Things

World Economic Forum. (2013). *Unlocking the Value of Personal Data: From Collection to Usage*. World Economic Forum.

World Resource Institute. (2014). *Global Forest Watch*. Retrieved from World Resource Institute: http://www.wri.org/our-work/project/global-forest-watch

World Wide Web Consortium (W3C). (2013). *Linked Data*. Retrieved from W3C: http://www.w3.org/standards/semanticweb/data

Zhu, P. (2013, December). *Business Intelligence vs. Big Data*. Retrieved from Future of CIO: http://futureofcio.blogspot.ca/2013/12/business-intelligence-vs-big-data.html