Technology Trends

Big Data

Enterprise Architecture, Chief Technology Officer Branch

Version 0.1
Date 2019-6-26

Shared Services Canada
Services partagés Canada

Canada

Table of Contents

Business Brief **3**
Technical Brief **3**
Industry Use **5**
Canadian Government Use **8**
Implications/Impact for Shared Services Canada (SSC) **8**
Value Proposition 8
Challenges 9
Considerations 10
Annex A – Big Data Further Defined **12**

Business Brief

Big Data is data of high variety, velocity, veracity, volume, and value that pushes limits of traditional tools and infrastructure, and demand cost-effective and innovative methods to process or extract value out of data. It is not a single technology but a combination of old and new technologies that helps organization gain actionable insight; and it is the capability to manage a huge volume of disparate data, at the right speed, and within the right time frame to allow real-time analysis and reaction.

Big data will be examined to uncover information including hidden patterns, unknown correlations, market trends and customer preferences, which is also called Big Data Analytics, to help organizations make informed business decisions, which in turn offers various business values, such as increasing productivity, reduce cost, faster results, improved operations, etc.

Big Data has been used broadly in industry fields, Sciences, Healthcare, banking, manufacture with combination of other technologies and data analytics.

It is important to note that Big Data is no longer included on the Gartner Hype Cycle[i]. Gartner sees it as a "matured" concept that is now very productive, and with all of the advancements made within this field, it is relatively simple to implement a big data solution within an organization. This is not to trivialize the complexity of a big data deployment, but it is just the present reality that big data is now another "tool in the box" to be used. In fact, big data is not an end in itself, it is considered to be a facilitator of other technologies, like artificial intelligence (AI) and machine learning, which requires large amounts of training data to become accurate.

Technical Brief

Big data is just a concept, it is not a technology in and of itself. Instead it facilitates and enables new methods of analysis that did not exist before the age of powerful computers and microchips. When it comes to using big data, organizations are only limited by computer capacity and the skill of a data analysis team. Almost any organization can have access to powerful solutions now that there are so many more tools available to process the variety of data. Big Data is considered a mature concept by Gartner because of the availability and sophistication of solutions available on the market. Real time, or live data can now also be analyzed to come up with valuable insights. This has changed the way data is handled, in the past there was a top down approach where data was used to support executive intuition, now analysts follow a bottom up approach where data is mined for insights that can lead strategic decisions

To make use of big data, organizations must shift their culture to be more data driven. Essentially, organizations should have a framework in place that efficiently lets data flow through necessary pre-processing phases: extraction, transformation, and loading. These steps are key; properly cleaned data from trusted sources will become more

valuable as it is easier to mine for actionable insights. It has been estimated that 80% of time and energy of a project is devoted to data preparation because most available data is raw, unstructured, and untagged.

From a technical standpoint, for data to be considered "Big Data", it must be a very large data set that could contain multiple types of digital data and large data volume that will be all stored in multiple distributed databases systems and distributed file systems. This necessitates the use of parallel computing to process the data simultaneously due to the data volume, disparity, and complexity of tasks. Many Big Data solutions are built around a parallel and distributed architecture called MapReduce, which was developed by Google. MapReduce allows distributed and parallel processing of queries on large data sets in a distributed environment. There are many variations on this core concept, and, at present, completely new ways of handling distributed data sets are also being developed, such as Big Data Cloudwhere data is stored and processes entirely in the cloud.

It is important to understand that simply having large amounts of data is useless if nothing is being done with it. Big Data is a concept that enables other technologies to function and produce valuable information. Data Mining is the process of discovering patterns in Big Data and it involves methods that combine machine learning, statistics, and database systems. Data Analysis on the other hand uses similar methods but one of its focus is on descriptive information. Currently, there is a marked shift in the handling of Big Data, from extensive human involvement in data manipulation to the use of AI, in handling common manipulation tasks. Big Data has actually allowed AI to advance at faster speeds because there is more variety and volume of data available for the training and testing of algorithms. Since machine learning algorithms of AI can continue to be optimized if they are constantly fed new data, they are iterative technologies that continue to be improved as time passes.

All Big Data analytics tasks fall into three main categories:[ii]

- **Descriptive Analytics:** this cluster of tasks deals with summarizing data to describe the way it behaves currently and in the past. The main goal of descriptive analytics is to find out why the data behave the way it does, and apply that knowledge to future decisions. It is estimated that more than 80% of business analytics are based off of descriptive processes.

- **Predictive Analytics:** these processes deal with predicting what is likely to happen based on available information. It is important to note that no forecast is perfect since they are probabilistic in nature as only the probability of occurrence can be extracted using these types of processes.

- **Prescriptive Analytics:** this is the next step in predictive analytics, where a course of action is prescribed by the processing model. Prescriptive analytics requires a

predictive model made up of two components: actionable data and a feedback system that monitors the outcome produced by the action taken. It allows users to assess what could potentially happen if certain parameters change, giving them an idea of what they should do to get desired results.

Industry Use

Big Data is considered to be a mature market, meaning there is a large amount of vendor solutions available for almost any perceived needs. These solutions can exist locally, meaning they are hosted within an organization's servers, they could be on the cloud, or they can be a mixed implementation where information is stored partially on the cloud and locally. According to a forecast done by Statista (a German online portal for statistics), the global market size for Big Data is currently at $49 billion United States Dollars (USD) and it is expected to reach $103 billion USD by 2027.[iii] Since Big Data solutions are widely adopted and many use cases have been studied and developed, this section provides a broad overview of how Big Data can be used.

Sciences

Big Data has changed the way many scientific fields operate. With the plethora of data available and the low cost of processing information, there has been a shift in how hypothesis are being tested; rather than a scientist coming up with a hypotheses and testing it, the data can now suggest hypothesis and relationships between data points.[iv] Scientists now have more ways to explore the vast amounts of data being generated by their experimentations.

The following examples are ways that big data is being used in the scientific community:

- Storage and processing of large amounts of experimental data. For example: Cern's Large Hadron Collider generates some 22 petabytes of data every year that is analyzed in a network of over 150 computing centres worldwide.[v]
- Improvements in monitoring methods: The monitoring of environmental, geographic, climate change information is no longer constrained by localized sampling. Big Data allows global data sets to be combined and analyzed, which gives scientists a better understanding of how the environment behaves.[vi]
- Resource management in agriculture: Agricultural organizations are crowdsourcing data from remote sensors and publicly available data sets to improve the land use efficiency of farmers. This can inform farmers of what types of crops they should plant and even the likelihood of their machines breaking down.[vii]

Healthcare

Patient data is being collected and anonymized in order to better understand the outcomes of patient treatments and how certain drugs will behave on specific diseases. This data, along with best practices, medical journals, and textbooks are also being fed into clinical decision support systems (CDSS) that are used to support doctors in recommending treatment plans. Some CDSSs are powered by AI, such as IBM Watson for Oncology.[viii] Other uses for big data in health care include:

- Pharmaceutical research and development
- Analyzing patient behavior
- Analyzing reimbursement claims and costs for fraud detection
- Early-warning analytics
- Drug cost-effectiveness
- Wearable technology to track individual health

Marketing and Business Intelligence

Big data allows managers and marketers alike to monitor in real time how their business efforts are performing. Through the following examples of tasks enabled by big data, organizations can better tailor their offerings and optimize their business strategies:

- Consumer behavior analysis
- Location based marketing
- Pricing optimization, dynamic pricing
- Market basket analysis
- Sentiment analysis
- Churn and survival rates
- Customer relationship management, loyalty management

Banking / Insurance / Securities

The banking industry is one of the earliest adopters of big data, which means that they have productive solutions that let them get a lot of value out of their data. The insurance and securities industries are closely related and benefit from some of the same advances in the sophistication of big data solutions. The following examples are use cases already in use within those industries:

- AI advisors for stock trading and investment strategies
- Automated Teller Machine (ATM) service optimization
- Credit risk assessment and credit scoring
- Fraud detection and money laundering detection
- Customer categorization and segmentation to determine optimal product offerings
- Algorithmic trading: taking in unstructured data through social media and news feeds to trigger automatic trades
- Disaster recovery costs

Manufacturing

Big data has also affected manufacturing processes and has made them more efficient. By using past data, organizations can optimize and fine tune their operations management strategies. Some use case examples include:

- Just in time supply chain optimization
- Product configuration planning
- Capacity allocation and supply network modeling
- Automated quality assurance
- Predictive maintenance
- Safety monitoring based off of sensor information
- Market pricing and planning based off of product quality, seasonality, demand, and other supply factors

Communications, Media and Entertainment

Just like other industries listed, the media and entertainment industries have been heavily impacted with the introduction of big data analytics. Social media platforms have given this industry a direct channel to "listen in" on what consumers are saying, which has changed the way entertainment is made and distributed. The following are some examples:

- Media recommendation algorithms (i.e. Netflix, Spotify, Youtube)
- Content optimization: determining what to make based off of consumer interests
- Content identification (i.e. Shazam, Youtube)
- Lifestyle monitoring for determining upcoming consumer trends in fashion

General Government Applications

Governments around the world have benefited from the relatively recent capability of being able to analyze large quantities of information. This has led to a new concept of analytics called civil analytics. Civil analytics is the use of citizen data to monitor government programs, as well as to come up with new initiatives. Data can also be used to discover insights that could change the direction of policy making. Programs could be measured in real time for their performance and precise changes can be made effectively. Other examples of the government use of Big Data include:

- Calamity management
- Crowd monitoring
- Preventative law enforcement and law enforcement efficiency
- Traffic optimization

Canadian Government Use

Currently, many departments within the Government of Canada (GC) are using or planning to use Big Data analytics to make more benefits from the data they already have. Additionally, some departments are offering analytical services on request basis. The National Research Council has a Digital Technologies Research Centre with a business unit focusing entirely on data analytics. They have a team of data scientists and machine learning experts that can be accessed on an on-demand basis. They provide services to "organize and analyze large data sets to discover patterns and trends, provide explanations and make predictions in order to create actionable knowledge from its analysis."[ix] Similarly, Innovation, Science and Economic Development Canada's (ISED) Big Data Analytics Centre visually explores the data it has collected about Canada's wireless spectrum.[x] In the same vein, the GC directly supplies research data of business intelligence capabilities for businesses on the Canada.ca portal.[xi]

The Canada Revenue Agency (CRA) has many initiatives to use the data it currently has to track and find taxpayers who are suspected of fraud. Using machine learning models, the CRA has been able to categorize taxpayers into different classes based on how likely they are to properly file and pay their taxes.[xii] Based on the taxpayer's category, the CRA has attempted different collection methods to see which one worked best. CRA is also mining the social media feeds of Canadians that it suspects of evading taxes and combines this information to create risk assessments.[xiii]

Another example of using Big Data within the public service is Agriculture and Agri-Food Canada (AAFC). Using data collected from across Canada, AAFC has been empowered to create public geomatics tools relating to the earth's surface. Big Data is once again an enabler this time for the creation of interactive maps that visualize crop trends and changes in land use.[xiv]

Implications/Impact for Shared Services Canada (SSC)

Value Proposition

Shared Services Canada (SSC) is already well positioned within the GC to implement and deliver on the promises of Big Data. An opportunity provided by the maturity of the market is that highly specialized knowledge is not always needed to use big data tools. Since many use cases have already been proven useful, many accessible solutions have been created by vendors to meet client needs. A proper deployment of a big data solution is a means to gain insights on how to improve an organization, and SSC can leverage this to refine strategic objectives and define clear goals. As it is already

being seen within SSC, data analytics can be used to measure and benchmark the efficiency of programs, which could be deployed at a larger scale across the GC. Since SSC is mandated to agglomerate GC data centers, it is also in a position to execute proper data handling strategies and database architecture from scratch so that data can be continuously used for years to come. SSC also has the opportunity to become the "Amazon Web Services" of the GC, meaning that unused data center computing space can be loaned out on an on-demand basis to partner departments when they need it.

As a GC service provider, it is essential that SSC focuses on its core services to strengthen storage, networks, communications, computing capabilities and the security of its infrastructure for a stable GC technology environment. Beside its mandated core operations, SSC could guide its future actions and offerings to seize the opportunities offered by Big Data in better supporting its partners and clients. These extended support opportunities need to be envisioned with consideration to SSC's areas of strengths and proper mitigation processes for any potential challenges they bring.

Transactional data is collected on a variety of issues every day by every GC organization. With the internet of things, the use of social media, sensor and videos for data collection, there is more than ever a vast amount of data collected every day. This data could be used over and over without its value being depleted. If this data is used in a coordinated manner, it will represent an asset for GC Big Data projects. By the same token, GC has a significant investment in digital infrastructure that could propel GC Big Data strategies.

Challenges

There are many challenges that present themselves with the implementation of a Big Data solution. In fact, before tools are implemented, the organization must have a clear process laid out to handle all of the preprocessing tasks. Those are actions that need to be performed on the data before the data can even be useful. The infrastructure to house big data is a continuous investment since the technology is continuously evolving and the data being generated is becoming more unstructured. There is also an organizational cultural change so that data will be respected and shared across entire organizations and people who need it will actually have access to it.

Privacy issues are a major concern inherent to Big Data. Sensitive data needs to be safeguarded and encrypted within a robust information technology (IT) security infrastructure, so as not to let nefarious parties have access to it. On this note, public trust in an institution can be easily lost when data breaches and leaks occur.

Additionally, the reliability of insights obtained through Big Data analytics can sometimes be suspicious due to the nature of the statistical methods used in analyzing data. There are two main concerns: first, Big Data analytics generally

focuses on correlation, not causation. Once a correlation is identified, often little or no effort is spent trying to understand why two variables are connected. Without such analysis, it is impossible to assess how reliable and cost-effective actions taken based on such correlations would prove to be over time. Second, databases used for analysis may not be randomized or representative of the population under study. For example, social media data may underrepresent the views of those with limited knowledge of or access to technology, such as seniors and low-income individuals. A compromised sample limits the ability to generalize the results to a broader context, which can lead to inaccurate decision making. These methodological shortcomings can be attenuated by using statistical techniques that search for causation, and by broadening samples to make them more representative.

Using machine learning techniques of AI also comes with its own challenges. These algorithms can be seen as a "black box", meaning that an external observer is unable to know how the algorithms arrive at specific conclusions. Depending on the type of algorithm used and the learning process that was chosen to train it, the algorithm may not be able to tell us exactly how it made decisions. Decisions made by Big Data enabled AI's will need to be monitored to ensure they do not learn to make the wrong decisions.

Considerations

SSC will need to consider the implications of Canada joining the Digital 9. This collaborative networks of countries with leading digital governments have a shared commitment to the open source development of future government systems and open government initiatives where partnered nations share open data.[xv] As the database hub for GC, SSC will potentially need to work within the proposed principles of digital development laid out within the organization's charter. Among Digital 9 nations, there is a trend towards the concept of the democratization of data, meaning that citizens are the owners of their own data and the government acts as a data steward. An example of a fully digital government is Estonia; citizens have access to all of their government held information (healthcare, taxes, banking, etc.) through the X-Road data platform, which links government databases together and retrieves information for a user in a single dashboard.[xvi] The GC and SSC will need to consider a similar approach that can be achieved if they are to stay technologically relevant.

With the upcoming digital governance directive and policy, there is a push for more interactive data that citizens can access. The GC will need to think about effective and flexible data architecture so that government departments can offer more interoperable digital services with a tell-us-once frame work. With tell-us-once concept that a citizen should only have to give his information once, instead of on an ongoing basis. Another consideration, with the push for open government data, is the subject of anonymizing information. If a citizen information is being collected, compiled, and used for projects, the information must be protected and anonymized before the data sets

can be published. The digital governance directive and policy also call for the use of open source software wherever possible, so this will also need to be considered when we adopt new technologies to deal with Big Data.

Another consideration is that of organizational change and change management. New technologies and strategies will be adopted to facilitate the use of Big Data enabled solutions. The way data is handled will also need to fundamentally change so that the right people have access to the right data at the right time.

Annex A – Big Data Further Defined

Volume: "Big Data" implies high volumes of data, high enough to cause challenges to managing, storing and treating of the data, since data in the digital universe grows exponentially.

Variety: What is called "Big Data" is data that goes beyond the conventional collections of data (structured or semi-structured data, such as surveys, administrative datasets, and operational/transactional data). Big Data includes a variety of data sources, and primarily many types of unstructured data: sensor data, text, video, audio, images, social media. Knowing that the majority of the data in the digital universe is unstructured (85 percent of all data, compared to 15 percent structured), and much of it is simply noise, new technological means of analysis are necessary to filter and classify digital data.

Velocity: The unstructured data is often produced at high velocity, such as through sensors or smart meters or even social network feeds, such as tweets. This unprecedented speed of data creation is seen as an opportunity for real-time analysis, especially in domains where timely alerts to a system are a necessity.

Veracity: Before being used for analysis, data is verified to ensure veracity, or truthfulness. Big Data comes in many formats and from disparate sources and, thus, the results of the analysis of the data require traceability and justification.

Value: Big Data is accompanied by a novel manner in which to decode data and extract value, a manner that has moved away from traditional statistics. The new field of "Big Data analytics" consists of new types of analysis, new ways of doing forecasting for detecting hidden patterns in the data that can lead to valuable insights.

References

Where is Big Data on the Hype Cycle?". Dan Bennet. Thomson Reuters. 26 September 2017. https://blogs.thomsonreuters.com/answerson/big-data-hype-cycle/

[ii] "Analytics: Predictive, Descriptive & Prescriptive". Big Data LDN. 17 October 2016 https://bigdataldn.com/analytics-predictive-descriptive-prescriptive/

[iii] "Forecast of Big Data market size based on revenue, from 2011 to 2027". Statisa https://www.statista.com/statistics/254266/global-big-data-market-forecast/

[iv] "How Big Data is Changing Science". Tom Chivers. Mosaic. 2 October 2018. https://phys.org/news/2018-10-big-science.html

[v] "Cern: Where the Big Bang meets Big Data". Nick Heath. Tech Republic. 22 May 2012. https://www.techrepublic.com/blog/european-technology/cern-where-the-big-bang-meets-big-data/

[vi] "Big Data: Explaining its uses in Environmental Sciences" Matthew Mason. https://www.environmentalscience.org/data-science-big-data

[vii] "Big Data: Explaining its uses in Environmental Sciences" Matthew Mason. https://www.environmentalscience.org/data-science-big-data

[viii] "What Watson for Oncology can do for your organization" IBM. https://www.ibm.com/ca-en/marketplace/ibm-watson-for-oncology

[ix] "Data Analytics Centre" National Research Council Canada. https://www.nrc-cnrc.gc.ca/eng/solutions/facilities/data_analytics.html

[x] "Communications Research Centre Canada" Science and Innovation. http://www.crc.gc.ca/eic/site/069.nsf/eng/h_00045.html#bigdata

[xi] "Research and Business Intelligence". Government of Canada. https://www.canada.ca/en/services/business/research.html

[xii] "Data Mining: Privacy Impact Assessment (PIA) Summary" Canada Revenue Agency. https://www.canada.ca/en/revenue-agency/services/about-canada-revenue-agency-cra/protecting-your-privacy/privacy-impact-assessment/data-mining-privacy-impact-assessment-summary.html

[xiii] "Canada Revenue Agency Monitoring Facebook, Twitter Posts of Some Canadians". Elizabeth Thompson. CBC News. 19 January 2017. https://www.cbc.ca/news/politics/taxes-cra-facebook-big-data-1.3941416

[xiv] "Geomatics: We're Got a Map For That". Agriculture and Agri-Food Canada. http://www.agr.gc.ca/eng/news/scientific-achievements-in-agriculture/geomatics-weve-got-a-map-for-that/?id=1492700093284

[xv] "Digital 5 Charter" December 2014. https://www.ict.govt.nz/assets/Uploads/D5Charter-signed-accessible.pdf

[xvi] "Estonia, the Digital Republic" Nathan Heller. 18 December 2017. https://www.newyorker.com/magazine/2017/12/18/estonia-the-digital-republic