CURRENT STUDIES IN COMPARATIVE EDUCATION, SCIENCE AND TECHNOLOGY, VOLUME 2, NUMBER 2, 2015, PP. 427-435

Big Data, the Cloud and Challenges of Operationalising Big Data Analytics

DOMINIQUE HEGER
DHTechnologies & Data Nubes, Texas, United States

JAMES OGUNLEYE
Middlesex University, United Kingdom

ABTRACT This paper highlights the current phenomenon of Big Data Analytics. The paper begins with a working definition of Big Data. It examines the application of Big Data analytics in Cloud computing, the main elements in operationalising Big Data analytics—people, tools and algorithms—as well as key challenges in operationalising Big Data analytics. The paper's conclusion is that users of analytics need to be clear-cut about why Big Data technologies are required and for what specific purpose. Another conclusion is that, the quality of available data and the modelling process are critical considerations in operationalising Big data analytics projects.

Keywords: Big Data, analytics, cloud, modeling process, data quality

Introduction

Defining Big Bata

In its simplest form, the term *Big Data* refers to the innovation that surrounds the possible uses of digital and physical information. The International Data Corporation [IDC] (2011) refers to Big Data as a new generation of technologies and architectures that are designed to extract value economically from very large volumes of a wide variety of data by enabling high velocity capture, discovery, and analysis features. Similarly, Big Data, as conceptualises in Oracle (2013, p.3) includes:

> *Traditional enterprise data—includes customer information from CRM systems, transactional ERP data, web store transac-*

> *tions, and general ledger data.*
>
> *Machine-generated /sensor data—includes Call Detail Records ("CDR"), weblogs, smart meters, manufacturing sensors, equipment logs (often referred to as digital exhaust), trading systems data.*
>
> *Social data—includes customer feedback streams, micro-blogging sites like Twitter, social media platforms like Facebook.*

The datasets referred to by the IDC (2011) and Oracle (2013) are largely non -metric data. They are huge and complex in volume, velocity, variety, veracity and variability and are significantly beyond the capability of standard data processing and analytic tools—and are threatening traditional computing architectures (see also McKinsey, 2011; Ogunleye, 2014).

Big Data Analytics

Big Data analytics is a process by which large sets of data, big data, are examined with a view to extracting actionable insights for strategic and operational decision making. The real value of Big Data lies is in the insights it generates when analyzed, the discovered patterns, the derived meaning, the indicators for decisions and ultimately the ability to respond to the business world in a timely fashion and with greater intelligence. Big data analytics refers to a set of advanced technologies that are designed to efficiently operate on large volumes (PetaBytes) of heterogeneous data. The technologies are based on sophisticated quantitative methods such as artificial intelligence, machine learning, neural networks, robotics, and computational mathematics and aid in exploring the data to discover unknown interrelationships and patterns. As already mentioned, Big Data analytics is moving from batch to real time. Intel conducted a survey of 200 IT managers in large enterprises in 2012 and discovered that while the ratio batch to real-time processing was basically 1:1, the trend is toward increasing real-time processing to two-thirds of total data management by 2015 (Intel, 2012). By today's standards, the statement can be made that the HW technology as well as the Big Data and Hadoop ecosystems are mature enough to support real-time Big Data analytics. Over the last couple of years, several Apache projects (such as Storm, Cassandra, HBase, Spark, Hama, or in-memory Hadoop from GRIDGain to name a few) are focused on providing (near) real-time support. In most scenarios, this is accomplished via some form of in-memory computing (IMC) feature. The IMC spectrum is vast, ranging from caching technologies embedded into Apache projects such as Spark or Hama to actual clus-

ter node HW setups that utilize some form of non-volatile memory technology (such as NAND flash, PCRAM, or RRAM) that do not require any disks (SSD or HD) in the cluster setup anymore. IMC aims at transforming the business. To illustrate, an application that is viewed by a company as a forecasting package and that runs overnight (for several hours as a batch job) is not a just a batch forecasting package anymore if the analysis can be completed within a few minutes (with Big Data, time-to-value is the key business driver). At that point the application becomes an interactive business tool that is changing the way a company is conducting business.

Big Data and the Cloud

As an IT infrastructure, organizations should evaluate and assess cloud computing as the underlying IT structure to support their Big Data projects. Most Big Data environments require a great number of clusters of servers to support the tools that process the large volumes, high velocity, and varied formats of Big Data projects. In a lot of companies, some form of IT Cloud is already deployed and hence can scale up or down as needed. Companies continue to store more and more data in Cloud environments, which represent an immense, valuable source/asset of information to mine. Further, Clouds do offer the scalable resources necessary to keep the Big Data costs under control. Cloud delivery models provide exceptional flexibility and value by being able to evaluate the best possible approach to each individual Big Data request Cisco (2012). To reiterate, investments in Big Data analysis can be significant and hence drive the need for an efficient and cost-effective IT infrastructure.

Private clouds offer that efficient, cost-effective model to implement Big Data analytics in-house, while potentially augmenting internal resources with public cloud services. Such a hybrid cloud option enables companies to use on-demand resources via public cloud services for certain analytics initiatives (such as short-term projects or proof of concept), and provide added capacity and scale as needed. Hence, Big Data projects may include internal and external sources. While companies often keep their most sensitive data in-house, huge volumes of data that may be owned by the organization or is generated by some third-party or public provider may be located somewhere externally (some may already be in a cloud environment). Moving that relevant external data behind a company's firewall can be a significant commitment of resources.

Analyzing the data where it resides, either in internal or public cloud setups often makes much more sense. Nevertheless, data services are needed to extract value from Big Data. Hence, depending on the requirements and the usage scenario, the best use of a company's IT budget may be to focus on Analytics as a Service (AaaS) that is supported by an internal private cloud, a

public cloud, or a hybrid model. The basic cloud service types for AaaS include the well known and widely available Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) models, respectively.

Operationalising Big Data Analytics

Big Data Analytics brings together management, information technology and modelling (Miller, 2014), and three elements are crucial in operationalising predictive analytics— e.g. are environment, models and architecture (Taylor, 2014). Take the first element. To ensure that the right issue is addressed or the right problem is solved, there is a need for an enabling environment that encourages team work and collaboration where everyone involved subscribe to a shared goal, agree to the problem that needed to be solved and, more importantly take ownership of the project. This is the foundation for a cost-effective, successful application of predictive analytics in any setting. The second element in operationalising Big Data analytics is model. The word 'model' in this context refers to the relation of one set of variables to another (Miller, 2014). Modeling is at the heart of predictive analytics project and it is extremely important that an organisation has in place a process for developing analytic models. According to Taylor (2014), the modelling process has to be 'repeatable, industrial-scale' to ensure effective development of 'dozens or even thousands' of required analytic models. The third element in operationalising Big Data analytics is a robust architecture, which is critical in the deployment and management of Big Data analytic models 'in production systems.' The deployment and management of Big Data analytic models are a collaborative effort that will involve a 'collaborative team of data modelers, data architects, scoring officers, and validation testers' (Chu, et al, 2007, p2).

Challenges of Big Data Analytics

The systematic approach towards data collection to enhance randomness in data sampling and reduce bias is not apparent in the collection of Big Data sets. Big Data sets do not naturally eliminate data bias. The collected data may still be incomplete and distorted, which in turn can lead to skewed conclusions. To illustrate, Twitter is commonly scrutinized for insights about user sentiments. There is an inherent problem there though as approximately 40% of Twitter's active user base is merely lurking but not contributing. Hence, the statement can be made that the actual tweets come from a certain type of user (vocal and participative in social media) and not from a true random sample.

It is paramount to understand that Big Data is not just about technology

(Big Data Insight Group, 2011). Big Data has to start as a business project that incorporates the right people and the appropriate business processes. Too many Big Data discussions solely revolve around the benefits of the technologies and how they aid companies in gaining a competitive advantage. This is a problem as Big Data adopters may miss the big picture by excluding or underestimating the importance of the people and business aspects involved here. Any company considering a Big Data project has to first evaluate the business cases, specify the goals and objectives, and stipulate the outcomes of the proposed Big Data initiatives. After the people and business impact and outcome is clearly understood, the IT capabilities and requirements can be evaluated. Developing a roadmap of how to achieve the desired business outcomes provides the organization with the understanding of what is required from a financial and organizational perspective.

Data science reflects the general analysis of data. The term refers to the comprehensive understanding of where the data comes from, what the data represents, and how to convert the data into information that drives the decisions making process. Data science encompasses statistics, hypothesis testing, predictive modeling, as well as understanding the impact of performing computations on the datasets. Data science basically consolidates these skills to provide a scientific discipline for the analysis of data. For any company interested in Big Data, data scientists (actual people) are needed. Gartner defines the data scientist as an individual responsible for modeling complex business problems, discovering business insights and identifying opportunities through the use of statistical, algorithmic, mining, and visualization techniques. In addition to advanced analytics skills, this individual is also proficient in integrating and preparing large, varied datasets, architecting specialized database and computing environments, and communicating the results.

Most advanced analytics projects involve identifying relationships across many datasets. Hence, the data scientist has to be able to integrate, validate, and if necessary cleanse the data (high quality datasets is the key here). A data scientist may or may not have specialized industry knowledge to aid in modeling the business problems. Locating and retaining professionals with this wide range of skills is a major challenge in itself and hence it is not a surprise to note that data scientists are currently in very short supply. A study conducted by McKinsey (2011) shows that by 2018, the US alone will face a shortage of approximately up to 190,000 engineers with deep analytical skills, as well as a shortage of approximately 1.5 million managers and analysts with the know-how to make effective decisions based on the results of any Big Data project. Today, locating the people capable of conducting the Big Data projects is considered the biggest issue.

Knowing what data to connect and the way to harness the data are the basic requirements before any form of data analytics can be done. In addition to data that is already within the confines of an organization, a majority of

the data could be outside the organization. To illustrate, such data may include social media feeds (Facebook, Twitter, or LinkedIn), geospatial (geographic location) data, news data, weather data, credit data, or retail information. Companies that own such data are realizing the value of the data and offer it for sale. To illustrate, Microsoft operates the Azure Marketplace that offers datasets such as employment data, demographic statistics, real estate information, or weather data. In regards to the different (internal and external) data sources, understanding the different data manipulation, integration, and preparation procedures is obviously paramount as sound datasets are representing the core of any deep analytics project. Traditional RDBMS solutions impose performance and flexibility restrictions on these advanced analytics activities due to extensive design, implementation, and data movement issues. Hence, NoSQL and IMC based solutions may be more appropriate to achieve the business goals and objectives of these Big Data projects.

Other challenges of operationalising Big Data analytics include:

Data Quality: Data is the lifeblood of analytics. It is important that organisations and their analytics teams have a deeper knowledge of the quality of their data as poor data could 'cause serious consequences for the efficiency of organizations' (Andreescu, et al., 2014, p.15).

Model and modelling: Model refers to a 'representation of the world, a rendering or description of reality, an attempt to relate one set of variables to another' (Miller, 2014, p. 2). Modelling, therefore, is a mathematical representation of an entity and very important in any predictive analytics project. Modelling can be a challenge if the modelling process is not well understood. As Taylor (2014) explains, the modelling process has to be 'repeatable, industrial-scale' to ensure effective development of 'dozens or even thousands' of required predictive analytic models—in order to search for 'meaningful relationship among models and representing those relationships in models' (Miller, 2014, p.2).

Return on investment: There is evidence that return of investment is as higher as 250% in predictive analytics projects (for example) compare to the 89% return of investment of projects that focused solely on accessing information and seeking internal gains in productivity, according to a survey by the International Data Corporation (Vesset and Harries, 2011). There is also evidence that many organisations deploy analytics projects with little or cognisance of the return on investment and those organisations that did have 'struggled to see a meaningful' ROI (Accenture, 2013).

Business case for Big Data analytics: There should be a business case for a Big Data analytics project. In other words, any decision about embarking on Big Data analytics project must reflect the business proposition and must

not be predicated on information technology infrastructure. This is what Heger (2014, p. 47) says about any organisation considering a big data analytics project, an argument that also applies to any predictive analytics project:

> *Any company considering a Big Data project has to first evaluate the business cases, specify the goals and objectives, and stipulate the outcomes of the proposed Big Data initiatives. After the people and business impact and outcome is clearly understood, the IT capabilities and requirements can be evaluated. Developing a roadmap of how to achieve the desired business outcomes provides the organization with the understanding of what is required from a financial and organizational perspective*.

Summary and conclusion

The term *Big Data* describes the innovation that surrounds the possible uses of digital and physical information. But, in using the information, users of analytics need to be clear about why Big Data technologies—architecture in particular—required and for what specific purpose. More so, the quality of available data, the modelling process, quality and integrity assurance of the analytics process are critical considerations in operationalising Big Data projects. The starting point in any analytics project is for champions of analytics to state in clear and compelling terms a business proposition and the kind/s of questions the organisation seeks answers to. People, tools and algorithms are at the heart of Big Data analytics and key to operationalising Big Data projects.

Correspondence
Dominique Heger
DHTechnologies & Data Nubes
10251 Twin Lake Loop Dripping Springs
TX, 78620 USA
Email: info@ dhtusa.com

References

Accenture (2013). Analytics in Action: Breakthroughs and Barriers on the Journey to ROI, online: http://www.accenture.com/sitecollectiondocuments/pdf/accentureanalytics-in-action-survey.pdf; accessed: 28.5.15.

Andreescu, I. A., Anda Belciu, A., Alexandra Florea, A. and Diaconita, V. (2014). Measuring Data Quality in Analytical Projects, Database Systems Journal, 5 (1), pp. 15- 25.

Big Data Insight Group (2011) The 1st Big Data Insight Group Industry Trends Report., 2011 [online] http://www.thebigdatainsightgroup.com/site/sites/default/files/The%201st% 20Big%20Data%20Insight%20Group% 20Industry%20Trends%20Report_0.pdf.

Chu, R. , Duling, D. and Thompson, W. (2007) Best Practices for Managing Predictive Models in a Production Environment, SAS Global Forum 2007, Paper 076-2007, SAS Institute Inc. Available: http://www2.sas.com/proceedings/forum2007/076- 2007.pdf; accessed: 20 September 2012.

Cisco (2012). The Internet of Things: How the Next Evolution of the Internet is Changing Everything, 2012, [online] http://www.cisco.com/web/about/ac79/docs/innov/ IoT_IBSG_0411FINAL.pdf.

Heger, D. (2014). Big Data Analytics—'Where to go from Here', International Journal of Develkopments in Big Data and Analytics, 1, 1, pp.42-58.

IDC (2011). The 2011 Digital Universe Study: Extracting Value from Chaos, 2011, [Online] http://uk.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf. IDC (2012) Worldwide Big Data Technology and Services 2012-2015 Forecast, 2012 [online].

Intel (2012). 'Big Data Analytics', Intel's IT Manager Survey on How Organizations Are Using Big Data, 2012 [online] http://www.intel.co.uk/content/dam/www/public/us/en/ documents/reports/data-insights-peer-research-report.pdf.

McKinsey (2011) Big Data: The Next Frontier for Innovation, Competition and Productivity, McKinsey Global Institute, available: file:///C:/Users/admin/Downloads/MGI_big_data_ exec_summary.pdf.

Miller, M. T. (2014). Modeling Techniques in Predictive Analytics: Business

Problems and Solutions with R, Pearson Education, Inc .

Oracle (2013). 'Oracle: Big Data for the Enterprise', White Paper June 2013, available: http://www.oracle.com/us/products/database/big-data-for-enterprise-519135.pdf; accessed: 29 Oct 2014.

Ogunleye, J. (2014). The Concepts of Predictive Analytics, International Journal of Developments in Big Data and Analytics, 1, 1, pp. 86-94

Taylor, J. (2014). 'Three steps to put Predictive Analytics to Work, *Decision Management Solutions*, SaS [online] https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper2/three-steps-put-predictive-analytics-work-105837.pdf; accessed: 20 August 2014.

Vesset, D. and Harries, D. H. (2011). 'The Business Value of Predictive Analytics', White Paper, International Data Corporation (IDC), Online: http://www.nexdimension.net/resources/products/ibm/spss/ibm-spss-predictive-analyticsbusiness-value-whitepaper.pdf; accessed: 3.5.15. Han, J., Kamber, M., and Pei, J. (2011) Data Mining.