International Journal of Knowledge, Innovation and Entrepreneurship
Volume 2 No. 3, 2014, pp. 38—54

Big Data Analytics—Where to go from Here

DOMINIQUE HEGER
DHTechnologies & Data Nubes, Texas, United States
Received 30 May 2014; received in revised form 9 November 2014; accepted 20 November 2014

ABSTRACT Big Data Analytics and Cloud Computing are headlining the current IT initiatives. The information pool that is generated worldwide doubles every 20 month. Echoing McKinsey (Manyika et al., 2001), Big Data refers to dataset sizes that are beyond the ability of typical database tools to capture, store, manage, and analyze the data assets. There is no explicit definition of how big a dataset has to be to be labeled Big Data. Nevertheless, new technologies have to be in place to manage the Big Data phenomenon. IDC 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 (IDC, 2011). Volume is synonymous with *big* in Big Data. Volume is a relative term though where smaller-sized companies may deal with TB of data as opposed to larger firms that may operate on PB and EB datasets. Anyhow, the data volume will continue to grow, regardless of an organization's size. Data comes from a variety of sources, both internal and external and in a variety of data types. So to reiterate, the term Big Data refers to the vast amount of heterogeneous (structured, semi-structured, unstructured) data that is generated at a rapid pace and basically yields the traditional data analysis tools used for RDBMS environments as useless. Further, the proliferation of smart-phones, tablets, sensors, and cameras combined with the popularity of social networks and the ease of world-wide Internet access contributes to the fact that most data generated today is labeled as unstructured or semi-structured, respectively. Traditionally, Hadoop (HDFS and MapReduce) was considered as the Java based, parallel/distributed batch processing platform of choice to analyze massive amounts of data (PetaBytes+). Since 2012 though, there is a clear shift in that organizations are moving the majority of their data management tasks from batch to real-time (Intel, 2012). The Big Data and Hadoop ecosystems as well as Cloud Computing continue to evolve. Organizations are moving beyond questions such as what and how to store towards deriving meaningful and timely analytics from their data to interactively respond to real business scenarios.

Keywords: Big Data, Analytics, Deep Learning, Cloud, Hadoop ecosystems

Big Data Analytics

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 cluster 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.

The statement can be made that real-time supports predictive analytics. Predictive analytics enables organizations to move to a future-oriented business perspective and represents the most exciting business opportunities for really deriving value from Big Data projects. While Business Intelligence (BI) focuses on the status quo and the past, answering questions such as *what was the sales volume for the last quarter in region x*, predictive analytics and data mining address issues such as *what will happen over the next 6 months if this trend continues*. Hence, real-time data provides the prospect for fast, flexible, and accurate predictive analytics that allows companies to quickly adapt to the ever changing business world. The faster the data is analyzed, the more timely the results and hence the greater the predictive value.

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. As discussed, 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.

Status Quo

The velocity of data in terms of the generating and delivery frequency is also a characteristic of big data. Conventional understanding of velocity considers how rapidly the data arrives, is stored and how swiftly it can be retrieved. In the context of Big Data, velocity has to be applied to data in motion (aka the speed at which the data is flowing) as well. The value of Big Data solutions is in converting the 3 V's (volume, velocity, variety) into tangible action items:

- Predicting probable future occurrences and determining the course of action that is most likely successful
- Analyzing the data in real time (or close to real time) and determining the appropriate actions
- Reviewing what happened and determining the next set of actions to be taken

In general, 3 current scenarios allow companies to leverage Big Data solutions:

- The technologies driving Big Data projects have matured to the point where actual implementations are practical and successful
- The underlying infrastructure cost to power the analysis has dropped considerably making it economical to deceiver the vast data and information pool
- World-wide competitive business pressure has increased to the point where most traditional data analysis strategies are only offering marginal benefits. Big Data has the potential to deliver the competitive edge necessary to advance a business in today's marketplace.

Over years, companies allocated transactional (structured) data and utilized batch processes to generate data sets that are stored in relational databases. The analysis of that data is retrospective and the evaluation of the datasets is based on past business operations. Over the last few years, new technologies enabled improvements in the data capture, storage, and analysis cycle. Most companies today capture different types of data from many sources such as emails, text messages, blogs, social media feeds, and audio/video files. Store and process options such as Hadoop MapReduce and/or In-Memory Computing (IMC) provide optimized capabilities for different business goals and objectives. IMC is feasible today due to the rather steep price-drop for memory modules (Gartner, 2011a). Some Big Data solutions combine DRAM with NAND flash memory arrays to form a cost-effective IMC solution. Data analysis can be done in real time (or close to real time), operating on entire datasets rather than on summarized subelements. Further, the proliferation of tools to interpret and analyze the data has increased significantly via visualization solutions such as IBM's ManyEyes, Lavastorm, or Plateau.

Current Market Evaluation

The swift adoption of mobile devices that are cheap, powerful, and equipped with applications and functionalities that almost resemble a traditional desktop system is considered a major driver of the continuing growth of unstructured data. Gartner estimated the 2012 smartphone shipment to be around 467.7 million units (Gartner, 2014). By 2015, the expected number of smartphones used will approach 1.1 billion. The market adoption rate of tablets is also expected to increase significantly over the next few years, further contributing to the growth of the world-wide datapool. In 2012, shipment of tablets reached 118.9 million units, a number that is projected to rise to 369.3 million units by 2015 (Gartner, 2012b). The market adoption of these mobile devices combined with the prevalence of mobile Internet access increases user connectivity via social media networks and for many reflects the communication platform of choice as well as the main source of information. This phenomenon provides a huge opportunity for companies to generate business opportunities by efficiently and effectively evaluating and analyzing this vast unstructured datapool. In 2010, Gartner

reported that more than 65 billion devices were connected to the internet. By 2020, Gartner estimated this number to increase to approximately 230 billion. These connected devices, ranging from smart meters to a wide variety of sensors and actuators continually send out vast amounts of data that has to be stored and analyzed. Companies that deploy sensor networks already adopted the required Big Data technologies to process the large amount of generated data.

Many of the technologies and tools powering the Big Data ecosystem are open sourced. To illustrate, the Hadoop framework, in conjunction with commodity HW (mainly Intel Core) and SW components such as Linux, R, Pig, Hive, Storm, and NoSQL solutions such as Cassandra, Mongo, CouchDB, Neo4j (graph data store) or HBase form the nucleus of many Big Data implementations today. The popularity and viability of these open source tools drove commercial vendors to introduce their own solutions (such as Oracle's NoSQL database) or to integrate these tools with their products (such as the EMC's Greenplum Community Edition that includes the open source tools Apache Hive, HBase, and ZooKeeper). It has to be pointed out that the traditional cluster or symmetric multiprocessing (SMP) solutions, while also providing a certain level of systems scalability are normally prohibitively expensive for many granular use case scenarios. Hence, the need for cost-efficient scalable HW and SW resources to process the related business logic and data volumes becomes paramount for SMB's as well as large business organizations alike. Most companies that are part of the technological evolution of Big Data are closely affiliated with the open source community. To illustrate, Cloudera, GridGain, Hortonworks, or MapR are all very active in various open source projects. The fact that open source technologies dominate the Big Data landscape is expected to continue as the technologies are swiftly changing and actual standards are not well established. On the flipside, this state posts a significant risk to any commercial vendor that intends investing in proprietary Big Data technologies.

Combining competent people with up-to-date (timely) information is the main differentiator in the business world, allowing for informed business decisions in an increasingly competitive landscape. In the past, market information was largely being made available by traditional market research companies and data specialists. Today, virtually any company (no matter the size) with an appropriate dataset can become a serious player in the new information age. The value of Big Data becomes more apparent to corporate leadership as companies are becoming much more data-driven (and data aware). According to O'Reilly, a data driven organization is one that acquires, processes, and leverages data in a timely fashion to create efficiencies, iterate on and develop new products and services to navigate the competitive landscape. There are neither any turn-key solutions nor comprehensive Big Data technology standards available today (no matter the claims from Oracle, EMC, or IBM based on their certification programs for their Hadoop based products). The main reason is the complex and diverse nature of every Big Data analytics projects (no 1 size fits all). Hadoop is considered today by many companies as being basically synonymous with Big Data as this Java based, distributed computing framework excels in processing vast amounts of unstructured data (it has to be pointed out though that the terms Hadoop and Big Data are not synonymous). The Hadoop Distributed File System (HDFS) provides a highly scalable, redundant data storage and processing solution that is used to execute a variety of large-

scale analytics projects. For large-scale structured data processing purposes, companies can use an open-source connector (such as Sqoop) to transfer data among their RDBMS solutions and HDFS. Further, Hadoop's MapReduce parallel processing capabilities have increased the extraction and transformation speed of data. Hadoop MapReduce can further be used as a data integration tool by reducing large amounts of data to a representative form that can be stored in a data warehouse. See (Heger, 2012) for a detailed discussion on the Hadoop ecosystem, HDFS, and MapReduce.

Some of the Big Data Challenges Today

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 (IDC, 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, vali-

date, 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.

The State of Big Data—Current, Around the Corner & Down the Road Technologies

Text Analytics

Text analytics reflects the process of deriving information from text sources. These text sources normally represent semi-structured data that may include web material, blogs, or social media postings (such as tweets). The technology supporting text analytics is based on linguistics, statistics, and machine learning (ML). Today's text analytics solutions use statistical models coupled with linguistic theories to capture patterns in human languages so that systems understand the meaning of the texts and can perform various text analytics tasks. These tasks can be as simple as entity extraction or more complex such as fact extraction or concept analysis. Entity extraction focuses on identifying an item, a person or any individual piece of information such as dates, companies, or countries. With Fact extraction, a fact is labeled a statement about something that exists, has happened, and is generally known. Fact extraction refers to identifying a role, a relationship, a cause or a property. Concept extraction functions identify an event, process, trend or a behavior.

In the near future, the consensus is that text analytics will develop into an essential data analysis approach for most companies as the attention shifts from analyzing structured to deciphering semi- and un-structured data. One major application of text analytics is the field of sentiment analysis where consumer feedback can be extracted from social media feeds, emails, and blog posts to improve customer relationships. One of the Hadoop components in the ecosystem that provides real-time analysis for these kind of tasks is Storm. Another area where text mining excels is the public security sector where text analytics can be used to scrutinize any text source for patterns that characterize potential criminal activities. The text analytics market is still evolving though. Language barriers (non English text), a variety of different text sources (some may be domain specific), as well as the requirement that knowing the metrics to use to determine the results have to be considered while evaluating this data analysis method.

Mobile Business Analytics

Mobile business analytics represents an emerging trend that is driven by the vastly growing attractiveness of mobile computing (Information Management (2013). For today's mobile business workforce, being able to remotely access the latest business insights generates the need for mobile business analytics. The field of mobile business analytics is decomposed into passive and active analytics (terms borrowed from operations research). Passive mobile business analytics is based on a push factor. Event-based alerts or reports can be pushed to the mobile devices after being refreshed at the backend. Passive mobile business analytics is not sufficient though to support the actual just-in-time analytical requirements of most business users. Hence, active mobile business analytics solutions enable the users to interact with the company-based business analytics systems on-the-fly. The solution may be implemented as a combination of push and pull techniques as a hybrid. To illustrate, the initial view of a business report could be a push process while any further analytical operations to obtain additional information based on the report could reflect a pull process. Mobile business analytics solutions provide powerful off-site decision making opportunities for stakeholders. In most cases, the decision makers only need access to a few key metrics and hence, providing up-to-date access to these metrics on mobile devices significantly improves the decision making process.

Predictive Analytics

Predictive analytics refers to a set of statistical and analytical techniques that are utilized to uncover relationships and patterns within large datasets to predict a potential behavior or some event (Gartner, 2012a). Predictive analytics may operate on structured and/or unstructured data as well as on data streams. The real value of predictive analytics is to provide proactive support systems that prevent any service impacting events from actually occurring. Gartner identifies 3 main methods for predictive analytics but the study outlines that in the future, a combination of the 3 discussed methods will be used in production environments.

1. The pattern-based approach compares real-time systems performance and configuration data with unstructured data sources that may include known failure profiles, historic failure records, or systems configuration data. Powerful correlation engines are required to identify statistical patterns within this vast, multifaceted and versatile data repository to determine if the current configuration and performance data indicates a likely failure scenario.
2. The rules-based approach focuses on a statistical analysis of the historic performance data, previously identified failure modes, and the results of stress and load testing (micro/macro) benchmarks to define a series of rules that are used as the comparison baseline against some real-time telemetry data. Each rule may cross-examine multiple telemetry (from the Greek *tele* = remote and *metron* = measure) data points, as well as other external factors such as the time of day or the environmental conditions. Based on the rules and the status quo behavior, various escalation routines can be triggered.
3. Statistical process control based models are based on control chart theory. Control charts represent the foundation for quality manufacturing processes and are regarded as an integral part to manage complex, process-driven systems. The combination of retrofit capable, real-time telemetry, improvements in data acquisition solutions, as well as the network capacity increase to support large data volumes implies that the statistical techniques used in the manufacturing area can now be utilized to industrialize (as an example) IT service delivery. Statistical anomalies can be identified and used to initiate appropriate actions to assure that service performance and functionality is at the appropriate, stable level.

As predictive analytics exploits patterns in historical and transactional data to identify future risk and opportunity scenarios, companies can acquire actionable insights and identify and respond to new opportunities much more swiftly. To illustrate, law enforcement agencies can identify patterns in criminal behavior and suspicious activities that aid in identifying possible motives and suspects, leading to more effective deployment of law enforcement personnel. Further, sensors combined with data analytics can be used to generate *heat maps* that define high-crime zones so that additional police resources can proactively be diverted.

Graph Analytics

Graph analytics refers to the study and analysis of data that can be transformed into a graph representation consisting of nodes and links. Graph analytics is a valuable option while solving problems that do not require processing all available data within a data set. A typical graph analytics problem requires graph traversal. Graph traversal describes the process of *walking* through the (directly) linked nodes. To illustrate, a graph analytics problem would be to identify by how many ways 2 members of a social network (such as Facebook or LinkedIn) are linked either directly or indirectly. Different forms of graph analytics are being used today.

- With a single path analysis approach, the objective is to identify a path through the graph while starting at a specific node. All the links and the corresponding

vertices that can be reached immediately from the starting node are evaluated first. Based on the identified vertices and a set of certain criteria, 1 is chosen and the first hop is executed. This process continues and the result will be a path consisting of a number of vertices and edges.

- With an optimal path analysis approach, the goal is to find the best path between 2 vertices. The best path could reflect the shortest path, the cheapest path, or the fastest path (depending on the properties of the vertices and the edges).
- With a vertex centrality analysis, the idea is to identify the centrality of a vertex based on several centrality assessment properties such as degree, closeness, or eigenvector. The degree centrality measure indicates how many edges a vertex has. The more edges, the higher the degree centrality. The closeness centrality measure identifies the vertex that has the smallest number of hops to other vertices. Hence, the closeness centrality of the node refers to the proximity of the vertex in reference to other vertices. The higher the closeness centrality, the more number of vertices that require short paths to the other vertices. The eigenvector centrality describes the importance of a vertex in a graph. Scores are assigned to vertices based on the principle that connections to high-scoring vertices contribute more to the score than equal connections to low-scoring vertices.

Graph analytics can be used in many areas. To illustrate, in the finance sector, graph analytics can be used to track and analysis money transfer pathways. A money transfer between bank accounts may require several intermediate accounts and graph analytics can be applied to determine the different relationships among the different account holders. Executing a graph analytics algorithm on financial transaction datasets aids in alerting banks to any possible cases of fraudulent transactions or money laundering. The usage of graph analytics in the logistics sector is widespread. Optimal path analysis is the obvious choice and graph analytics is hence used in most logistics distribution and shipping environments. One of the Hadoop components in the ecosystem that aids in graph analysis is the Neo4j graph data store.

In-Memory Computing (IMC)

IMC basically represents an analytics layer where detailed data (today up to TB's) is stored directly in systems memory (Pezzine, 2011; Pezzine, et al., 2012; Gartner, 2011b). The data may be from a variety of sources, and IMC is basically deployed to support fast (memory speed) query, analysis, calculation, and reporting functions (no disk or SAN IO is involved). Most IMC solutions are based on a combination of DRAM and NAND flash memory arrays. Disks are only used for backup or logging purposes. With traditional (RDBMS) disk and SAN based analytics platforms, the metadata has to be created prior to the actual analytics process. Examples of RDBMS metadata include (1) tables of all the tables in the database, their names, sizes and number of rows in each table or (2) tables of columns in each database, what tables they are used in and the type of data stored in each column. The metadata is basically modeled based on the analytical requirements. Modifying the metadata model to fulfill new requirements is rather complex and time consuming. IMC eliminates the need to pre-model the metadata for

every business requirement, significantly increasing the flexibility and performance of any data analysis project. To illustrate, the data can be analyzed as soon as it is available in memory. The speed advantage of IMC allows for a (near) real-time analysis of the datasets and powerful, interactive data visualization tools (such as Plateau or Lavastorm) can be used to further aid in the decision making process.

Although the average price per GB of DRAM is steadily coming down, RAM is still more expensive than the average price per GB of disk storage (Gartner, 2011a). Hence, some IMC solutions combine DRAM with NAND flash arrays to bring the price point down. Studies conducted by SAP have shown though that the TCO for an IMC solution can be lower than for a SAN based analytics platform. To reiterate, advancements in memory technology, particularly through the form of non-volatile memory, have changed the way data is being accessed. Non-volatile memory expedites the data access process and moves storage closer to the processing units. Today's NAND flash memory technology with persistent storage provides access latencies that vastly outperforms any disk based solution—including SSD configurations. Next to NAND flash memory, phase change memory (PCR or PCRAM) is advertized as the fast, low-power, random-access, non-volatile and potentially low-cost memory alternative to NAND. PCRAM uses a chalcogenide-based alloy that has formerly been used in other re-writable optical media solutions such as CD's and has several significant advantages over current NAND flash memory. First, it is more robust than NAND, as it has a much better lifespan (write cycles) compared to NANDs. Hence, PCRAM opens the door for developing simpler wear leveling algorithms that generate less HW/SW overhead. Second, PCRAM is byte-addressable and hence much more efficient than NAND while accessing smaller chunks of the dataset. From an actual implementation perspective, it is worth noting that Viking Technology[1] is shipping a memory board that combines DRAM and NAND flash memory to create a non-volatile standard DIMM card that can be used in servers. Another company, Crossbar[2] announces the first RRAM that could replace flash that stores up to 1 TB on a single chip.

From a SW perspective, existing Hadoop clusters can utilize In-Memory Hadoop solutions from GRIDGain, use Hadoop projects such as Storm or Cassandra, or configure ML solutions such as Spark, Hama, or HaLoop that provide IMC functionalities without the need to acquire the still rather expensive HW memory solutions discussed above.

Complex Event Processing (CEP)

A complex event reflects an abstraction of other base events and represents the collective significance of these events. CEP combines data from multiple sources to determine trends and patterns based on seemingly unrelated events. CEP represents a style of computing that is implemented by event-driven, continuous intelligence systems. A CEP system uses algorithms and rules to process streams of event data that is received from 1 to *n* sources to generate insights. CEP processes complex events, putting them in context to identify threat or opportunity situations (Big Data Insight group, 2011). To illustrate, sales managers receiving an alert message such as *today's sales volume is 25% above average* understand the magnitude of the situation much more swiftly than if

1. www.vikingtechnology.com
2.www.crossbar-inc.com

they were shown the thousands of individual sales transaction records (the base events) that contribute to the complex event. This information can then be used to guide an appropriate action as sense-and-respond business activities. The CEP computation process is triggered while receiving event data. CEP systems store large amount of events (in memory) and perpetually process the data to act accordingly while more event data arrives.

The CEP technology revolves around sense-and-respond and situation awareness applications. CEP can be used to analyze events within a network, power grid, or any other large (distributed) system to determine whether the environment is performing optimally, is experiencing some problems or has become the target of an attack. CEP can also be used for optimization analysis projects to aid activities such as inventory tracking, detecting homeland security threads, or supply chain management. With CEP, financial trading institutions can correlate worldwide stock, commodity, and other market movements to recognize potential opportunities or issues with their current portfolio holdings. CEP could also be used to examine internal corporate activities to determine if the company conforms to corporate policies or government regulations. With CEP, businesses can map discrete events to expected outcomes and relate them to key performance indicators, extracting insights that will enable them to allocate resources to maximize opportunity and mitigate risk. CEP solutions (from companies such as Oracle, Esper, Streambase, or Progress) are still very expensive and hence, CEP platforms can be considered an over-investment for many event processing applications that only process modest event volumes and simple processing rules. In most cases, event processing can be accomplished without purchasing a CEP solution. By embedding custom CEP logic into the existing business application framework, the outcome could be equally desirable at a much lower cost point.

Quantum Computing

While the current distributed computing infrastructure (including the Hadoop ecosystem) is considered as being very powerful by today's standards, to actually increase the computing power of these systems to address the ever-growing Big Data project requirements, it is necessary to add additional transistors into each cluster node (IDC, 2012). This effort is getting more and more difficult though as it is already a daunting task to add additional transistors based on the currently available technology. By 2012, the highest transistor count in a commercially available CPU is over 2.5 billion transistors (Intel 10-Core Westmere-EX). Ergo, a paradigm shift is necessary to meet future computing demands. Quantum computing refers to the fusion of quantum physics and computer science and represents a paradigm shift where data is represented by *qubits*. Unlike conventional systems that contain many small transistors that can either be turned on or off to represent 0 or 1, each qubit in quantum computing can be in a state 0, a state 1, or in a mixed state where it reflects 0 and 1 at the same time. This property is known as superposition in quantum mechanics (*Schrödinger's cat* comes to mind) and it provides quantum computers the ability to compute at a speed that is exponentially faster than conventional computers. For certain problems such as database search operations or factoring very large numbers (which is the basis for today's encryption

techniques), quantum computers can produce an answer many orders of magnitude faster than today's fastest systems (some problems today just cannot be solved within a reasonable time-frame). In quantum computing, a data value held in a qubit has a strong correlation with other qubits even if they are physically separated (Bell theorem—Spooky action at a distance). This phenomenon, known as entanglement, allows scientists to state the value of one qubit just by knowing the state of another qubit. Qubits are highly susceptible to any effects of external noise. Hence, in order to guarantee accuracy in quantum computing, qubits must be linked and placed in an enclosed quantum environment, shielding them from any noise.

Currently, quantum computing is still considered an (advanced) research project. Nevertheless, quantum computing receives significant funding and extensive research is being conducted on the HW as well as the SW (algorithm design) side and substantial advances have been made lately. Google has been using a quantum computing system (from D-Wave) since 2009 to research highly efficient ways to search for images based on an improved quantum algorithm discovered by researchers at MIT. In 2012, IBM research (based on work by R. Schoelkopf, Yale) derived a qubit that lasted for 1-10,000th of a second. This amazing achievement aids in lengthening the time-frame for error correction algorithms (to detect and resolve any possible mistakes). Generating any reliable results from quantum computing requires a very low error rate and hence, the IBM achievement is considered a breakthrough.

Deep Learning

A typical machine learning (ML) problem scenario is that a system is faced with a large set of data and on its own, is tasked to sort the elements of the data set into categories. A good analogy would be to ask a child to sort a set of toys without providing any specific instructions. The child may sort the toys by color, by shape, by function, or by something complete else. The ML scenario described here attempts the same thing just on a much larger scale. A system may be fed millions of handwritten digits and is supposed to guesses which digits look comparable, basically clustering the digits together based on similarity. The essential deep learning novelty is to design and implement models that learn these categories incrementally by trying to identify first lower-level categories (like letters) prior to attempting to obtain higher-level categories (like words).

So in a nutshell, the term deep learning describes a particular (novel) approach to implementing and training artificial neural networks (ANN). ANN's were introduced in the 1950s and so they have been around for a long time. Artificial Intelligence in general has been labeled as an amazingly promising research idea, but the practical deployment did not really happen until a couple of years ago. Simplified, a neural network can be described as a decision-making black box system. ANN's consume an array of values (that may represent pixels, audio waveforms, or strings), execute a series of functions on the array and output 1 or more values as the result. The output normally reflects a prediction of some properties one is trying to estimate based on the input (to follow one of the initial deep learning projects conducted by Google, to determine if an image is a picture of a cat). The functions that execute inside the black box are con-

trolled by (1) the memory of the neural network and (2) the arrays of numbers that are labeled as the weights that define how the inputs are combined and recombined to generate the results. Processing real-world problems such as Google's cat-detection project requires very complex functions, which implies that the arrays are rather large (sometimes in the range of millions of values/numbers). The biggest obstacle to really using ANN's has been to determine how to set all these gigantic arrays to values that efficiently/effectively transform the input signals into output predictions.

One of the goals and objectives of ANN research is that the networks should be *teachable*. It is rather trivial (at a small scale) to demonstrate how to feed a series of input examples and expected outputs into the system and to execute a process to transform the weights from initial random values to progressively better numbers that produce more accurate predictions. The main problem is how to do the same thing at a large scale while operating on complex problems (such as speech recognition) where a much larger numbers of weights is involved. In 2012, Krizhevsky et al. (2012) published a paper outlining different ways of accelerating the learning process, including convolutional networks, smart ways to utilize GPUs, as well as several novel mathematical approaches such as Rectified Linear Units (ReLU) and dropout procedures.

The term *dropout* refers to a technique that avoids overfitting in neural networks. The researchers showed that within a few weeks, they were able to train a very complex network at a level that outperformed conventional approaches used to solve computer vision related tasks. The new techniques outlined by the researches have also been successfully applied to natural language processing and speech recognition related projects and so form the heart of deep learning. With most machine learning projects, the main challenge is in identifying the features in the raw input data set. Deep learning aims at removing that manual step by relying on the training process to discover the most useful patterns across the input examples. It is still required though to design the internal layout of the networks prior to the training phase, but the automatic feature discovery step significantly simplifies the overall process. Deep learning excels at these *unsupervised learning* tasks but still has much room for improvement. To illustrate, the much hyped Google system that learned to recognize cats outperformed its predecessor by about 70%, but still only recognized less than 1/6th of the objects on which it was trained. Rationally, deep learning is only part of the equation of building intelligent machines. Today's techniques do not allow representing causal relationships, perform logical inferences, or integrate abstract knowledge. The most powerful artificial intelligence systems today use deep learning as 1 element and fuse it (in a rather complex manner) with other techniques ranging from Bayesian inference to deductive reasoning.

Ambient Intelligence

Addressing the question of 'what is the next big thing when it comes to Big Data' from a different angle leads to a discussion on Ambient intelligence (AmI). Ambient intelligence reflects a fast emerging (research) discipline that brings intelligence to everyday environments that are sensitive to people. In other words, AmI's vision is that people are living easily in digital environments in which the electronics are sensitive to the

people's needs, personalized to their requirements, anticipatory of their behavior and responsive to their presence. Ambient intelligence research builds upon advances in sensors and sensor networks, pervasive computing, and artificial intelligence (AI). So from an AI perspective, the idea is to hide the complexity of the technology from the developers by providing intelligent tools that act as the building blocks for Big Data solutions.

Concluding Remarks

The proliferation of the world-wide data volume over the last few years has been staggering. Converting this huge asset (data) into tangible business action items is for most companies today a daunting task (for some completely out of reach). The emergence of the Big Data and Hadoop ecosystems is only the beginning and reflects the first step in dealing with the large datasets that are perpetually/rapidly growing and consist of many different data types. From a technical perspective, many further improvements in IMC, ML, Deep Learning, or CEP (down the road, quantum computing will help here as well) are necessary to successfully address the Big Data business vision of *extracting value for a company* or in other words, to provide companies with business results in a timely fashion so that the stakeholders can make informed decisions to advance the business. This is and will not be possible without data scientists and managers that can actually make appropriate decisions based on the results of any Big Data study. With most Big Data projects, the trend is towards real-time processing and hence, predictive analytics. The advanced technologies discussed in this paper, combined with the ever evolving Hadoop ecosystem where many projects these days focus on real-time processing fuel the movement. Most Big Data projects do require a substantial amount of server systems, hence any form of Cloud Computing should be assessed as the potential delivery environment for AaaS. The Cloud does provide the flexibility, elasticity, and cost incentive for many companies to further explore the possibilities of engaging Big Data projects into their business cycle.

Correspondence

Dominique Heger, PhD
DHTechnologies & Data Nubes
10251 Twin Lake Loop
Dripping Springs, TX, 78620
USA.
Email: info@ dhtusa.com
Phone: +1(512) 773 1938

References

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.

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.

Gartner (2011a) Weekly Memory Pricing Index 3Q11 Update, 2011 [online] http://www.gartner.com/technology/core/products/research/markets/hardwareStorage.jsp;

Gartner (2011b) Emerging Technology Analysis: The Future and Opportunities for Next-Generation Memory, 2011 [Online] https://www.gartner.com/doc/1832620/emerging-technology-analysis-future-opportunities.

Gartner (2012a) Big Data, Bigger Opportunities: Investing in Information and Analytics, 2012 [online] http://www.gartner.com/technology/research/big-data/.

Gartner (2012b) Worldwide IT Spending Forecast, 3Q12 Update, 2012 [Online] https://www.gartner.com/doc/2238215/hightech-tuesday-webinar-gartner-worldwide.

Gartner (2012c) Worldwide Media Tablets Sales to Reach 119 Million Units in 2012, 2011, [online] http://www.gartner.com/newsroom/id/1980115.

Gartner (2014) Gartner Says Annual Smartphone Sales Surpassed Sales of Feature Phones for the First Time in 2013, [Online] http://www.gartner.com/newsroom/id/2665715.

Heger, D. (2012) Hadoop Design, Architecture & MapReduce Performance, CMG Journal of Computer Resource Management, available: http://www.cmg.org/publications/cmg-journal/.

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] http://www.idc.com/research/viewtoc.jsp?containerId=233485.

IDC (2011) IDC's Worldwide Big Data Taxonomy, 2011 [online] http://www.marketresearch.com/IDC-v2477/IDC-Worldwide-Big-Data-Taxonomy-6708183/.

Information Management (2013) Mobile Business Intelligence for Intelligent Businesses, 2013 [online] http://www.intelligenthq.com/technology/mobile-business-intelligence-market-on-the-rise/.

Intel (2012) 'Big Data Analytics', Intel's IT Manager Survey on How Organizations Are Using

Krizhevsky, S. Sutskever, I. and Hinton, G.E. (2012) "ImageNet Classification with Deep Convolutional Neural Networks", University of Toronto, [online] http://www.cs.toronto.edu/~fritz/absps/imagenet.pdf.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Byers, A. H. (2011) "Big data: The next frontier for innovation, competition, and productivity", McKinsey Global Institute, [online] www.mckinsey.com/mgi.

Pezzini, M. (2011) Net IT Out: In-Memory Computing, Thinking the Unthinkable Applications, 2011 [online] http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf.

Pezzini, M., Claunch, C. and Unsworth, J. (2012) Top 10 Strategic Technology Trends: In-memory Computing, 2012. [Online] https://www.gartner.com/doc/1914414/top--strategic-technology-trends.