Chapter 9
Custom Hardware Versus Cloud Computing in Big Data

Gaye Lightbody, Fiona Browne, and Valeriia Haberland

Abstract The computational and data handling challenges in big data are immense yet a market is steadily growing traditionally supported by technologies such as Hadoop for management and processing of huge and unstructured datasets. With this ever increasing deluge of data we now need the algorithms, tools and computing infrastructure to handle the extremely computationally intense data analytics, looking for patterns and information pertinent to creating a market edge for a range of applications. Cloud computing has provided opportunities for scalable high-performance solutions without the initial outlay of developing and creating the core infrastructure. One vendor in particular, Amazon Web Services, has been leading this field. However, other solutions exist to take on the computational load of big data analytics. This chapter provides an overview of the extent of applications in which big data analytics is used. Then an overview is given of some of the high-performance computing options that are available, ranging from multiple Central Processing Unit (CPU) setups, Graphical Processing Units (GPUs), Field Programmable Gate Arrays (FPGAs) and cloud solutions. The chapter concludes by looking at some of the state of the art solutions for deep learning platforms in which custom hardware such as FPGAs and Application Specific Integrated Circuits (ASICs) are used within a cloud platform for key computational bottlenecks.

9.1 Introduction

The exponential growth in technology has fuelled the rise of complex computing applications churning out reams of data and information which in turn needs to be processed using high-performance computing solutions, stored using mammoth

G. Lightbody (✉) • F. Browne
School of Computing and Mathematics, Ulster University, Shore Road, Newtownabbey, Co. Antrim, BT37 0QB, UK
e-mail: g.lightbody@ulster.ac.uk; f.browne@ulster.ac.uk

V. Haberland
Tungsten Centre for Intelligent Data Analytics, Goldsmiths, University of London, New Cross, SE14 6NW, London, UK
e-mail: v.haberland@gold.ac.uk

A.J. Schuster (ed.), *Understanding Information*, Advanced Information and Knowledge Processing, DOI 10.1007/978-3-319-59090-5_9

data centers and managed through the support of refined data governance. Such applications span a broad range of areas and disciplines and this spread is accelerating at a phenomenal rate. The world around us offers endless possibilities of monitoring and gathering data. Our cities, homes and even ourselves are amassed with technology for monitoring, collating and analyzing data. From a vision of smart cities (Townsend 2014) in which the very control of home heating is managed through analytical decisions (ODwyer et al. 2016) through to effective control of power generation (BritishGas 2017), it is clear to see how such analysis opens up the potential for affecting power consumption and ultimately impacts the global fuel crisis.

Through the development of powerful technology such as smart phones, wearable tech and sensors, we are now generating huge amounts of personal data on our daily lives, behavior, health and well-being. We are currently amidst a self-quantification era in which we wear sensors to report back on activity, behavior and well-being.[1] From a non-clinical aspect this enables a tracking of fitness and personal goals with the added dimension of social support through disseminating our personal metric data through social media communities. The direction this is going, is to a more biological level in that we are prepared to share biosignal metrics and signals such as Electroencephalography (EEG) (Terrell 2015) and even our own DNA (AncestryDNA™ 2016; 23andMe 2015) in the concerted goal to furthering ourselves and medical science.[2]

Continuing the discussion in the medical domain, a further source of high volume heterogeneous data is with digital records. Such an encompassing term spans far beyond text-based information to mammoth digital files of *x*-ray images, Magnetic Resonance Imaging (MRI) scans, recordings of EEG and possible Exome or Genome sequences. The image processing required for digital capture again needs to be of a significant quality as not to lose vital information from the record. Furthermore, methods to analyze and quantify what the images are showing indicate a necessity for high-performance computing solutions (Wang et al. 2010).

The result of such generation of huge volumes of data is referred to as big data. However, it is not only the sheer quantity of data created that defines big data, but there are also the four 'V's' (Hashem et al. 2015) that are recognized characteristics:

1. Volume: refers to the sheer amount of data coming from multiple resources.
2. Variety: refers to the heterogeneous nature of the data. That is data of the different types coming from the different collection mechanisms, such as sensors, physiological recordings, speech, video, text, social networks, to name just a few. In addition to the sheer amount of data, a major hurdle is in handling the diversity in data format and whether the data is structured or unstructured.
3. Velocity: refers to the speed at which the data is created and transferred.

[1]Quantified Self. http://quantifiedself.com/. Accessed: 2017-02-03.

[2]IGSR: The International Genome Sample Resource. http://www.internationalgenome.org/home. Accessed: 2017-02-03.

4. Value: The benefit of meeting such a challenge is the potential that by gathering such a diverse and large set of data then previously hidden trends and patterns can emerge through analysis.

Big data opens up a range of challenges along every stage of data handling, processing and analysis (Chen et al. 2014). The computational challenges are extreme and as such a range of solutions exists, where each platform is heralding scalability and performance advantages. In this chapter a high-level review is given of the range of common applications in which big data now features. The overview provides some insight into different solutions or examples of how the computational challenges have been met in these applications. A summary is provided of high-performance platforms, ranging from multiple CPU setups, GPUs, FPGAs and cloud solutions. The chapter concludes with a discussion around custom hardware solutions versus scalable on-demand cloud computing solutions, asking the question whether cloud computing holds all the cards? A peek into current technology trends is given suggesting that custom devices may be the support engine for computational enhancements for the cloud, while providing customers with the scalable and on-demand service that they require.

9.2 Applications

The range of applications involving big data is comprehensive and diverse, playing a role in personalized medicine, genomics, self-quantification through to monitoring financial markets or transactions. Smart cities and the Internet of Things (IOT) create a wealth of recordable data from the devices in homes through to cities. This section provides a high-level overview of some of the current big data challenges.

9.2.1 Genomics and Proteomics

In the last decade there has been a seismic shift in the technological advances for sequencing DNA. Edward Sanger developed the Sanger approach in 1975 using capillary electrophoresis and for decades this approach has been the technique employed. It is expensive and slow, limiting the opportunities for use. However, recent technological advances in sequencing has led to it being possible to sequence a whole human genome using a single instrument in 26 h (Miller et al. 2015). The enabler for this has been the development of High-Throughput Sequencing (HTS) which provides massively parallel sequencing power at an accelerated rate yet with significant cost reductions (Baker 2010; ODriscoll et al. 2013).

The reduction in costs has made HTS technologies much more accessible to labs and has facilitated their use in a broad range of applications and experimentation, including diagnostic testing for hereditary disorders, high-throughput polymorphism detections, comparative genomics, transcriptome analysis and ther-

apeutic decision-making for somatic cancers (Van Dijk et al. 2014). A review and comparison of sequencing technologies can be found in Metzker (2009) and Loman et al. (2012).

However, HTS generates enormous datasets, with the possibility of producing >100 gigabases (Gb) of reads in a day (Naccache et al. 2014). For these reasons, coupled with the challenges of integrating heterogeneous datasets, HTS sequencing data can be characterised as big data, and as such there lies a significant computational challenge. High-performance, cloud and grid computing are aspects of computing that have become ubiquitous with processing and analysis of HTS data (Lightbody et al. 2016), generated at ever increasing momentum. As the technologies are ever developing, sequencing could become a routine facet of personalized medicine (Erlich 2015).

9.2.2 Digital Pathology

Traditional microscopy involves the analysis of a sample, for example, a biopsy on a glass slide using a microscope. The domain of virtual microscopy has moved from viewing of glass slides to viewing of diagnostic quality digital images using specialised software. These slides can be viewed on-line through a browser or as recently demonstrated via a mobile device whereby the computational power of mobile devices provide a cost-effective mobile-phone-based multimodal microscopy tool which combines molecular assays and portable optical imaging enabling on-site diagnostics (Kuhnemund et al. 2017). Where more extensive computational power is required, some service providers have opted for cloud based virtual microscopy solutions which offer the promise of in-depth image processing of the tissue samples (Wang et al. 2010).

The drive towards personalized medicine has led to a deluge of personal data from heterogeneous sources. This big data challenge is discussed by Li et al. (2016), in which they highlight that "integrative analysis of this rich clinical, pathological, molecular and imaging data represents one of the greatest bottlenecks in biomarker discovery research in cancer and other diseases". They have developed a framework, Pathology Integromics in Cancer (PICan), to accelerate and support data collation and analysis. This framework connects the tissue analysis to other genomic information, enabling a full and comprehensive understanding to be attained.

9.2.3 Self-Quantification

We are in an era in which society is 'comfortable' with every aspect of their behavior and person being monitored and analyzed. Part of this, has been the birth of a Quantified Self (QS) movement in which the person collates data on their daily life and physiology. It is reported as "self-knowledge through numbers".[1]

The goal of such monitoring is often for self-improvement, whether it is to encourage more physical activity or to improve on lifestyle choices (Almalki et al. 2013). Alternately, it could come from the belief that by gathering enough data from enough people, then trends in the data can be found. This offers the opportunity to impact society's health and well-being, and not just benefit the individual.

The advances in personal devices such as smart phones and sensor technology have promoted the gathering of such vast resources of personal data, which can fall into the category of big data, due to the sheer amount of data, the heterogeneous nature of the data and the speed at which it needs to be processed and managed.

An emerging addition to the QS movement is in collecting and analyzing electrical activity of the brain. Measured using the EEG, evaluation and classification of brain function such as sensory, motor and cognitive processes can be made. With the advancements in electronics,[3] wearable sensors, algorithms and software development kits there has been a shift towards exploring other possible applications in which EEG can play its part. One organization[4] has developed a neuroscience platform to encourage users to perform "routine brain health monitoring". By many users sharing their EEG, it is envisaged that it may be possible to derive critical insight into brain health and disease.

As QS applications evolve, it is expected that advanced machine learning and pattern recognition techniques will be involved in the analysis of data coming from multiple heterogeneous sources such as wearable electronics, biosensors, mobile phones, genomic data, and cloud-based services (Swan 2013).

9.2.4 Surveillance

Surveillance, specifically videos, are becoming ubiquitous in a number of situations for the monitoring of activity. With threats of terrorism, crime events, traffic incidents and governance, we have seen a rise of surveillance across global cites. Alongside this increase, we have seen progress on research in the area of computer vision, whereby processing and understanding surveillance videos can be performed automatically and key tasks such as people segmentation, tracking moving entities, as well as classification of human activities have been undertaken. Big data and the four 'V's' are relevant to the surveillance domain due to the scope and volume of video data captured (Xu et al. 2016). It has been estimated by the British Security Industry Association that there are between 4 and 5.9 million cameras in the UK. A single camera can capture up to 48 GB of high-definition video a day. This results in issues with local storage through to the fusion of data from multiple video streams which may differ in terms of format. These issues lead to the processing of video analytics which has an impact upon terrorist prediction and governance. To address such needs, research has been performed in the area. This includes the study by Xu

[3]EMOTIV. https://www.emotiv.com/. Accessed: 2017-02-03.

[4]BrainWaveBank. https://www.brainwavebank.com/. Accessed: 2017-02-03.

et al. (2015) whereby a semantic based model called Video Structural Description was proposed to represent and organize video resources (Najafabadi et al. 2015).

Another application in the area has been work performed by Krizhevsky et al. (2012) where deep convolutional neural networks were applied to classify 1.2 million images in the ImageNet dataset, achieving top-1 and top-5 and error rates of 37.5% and 17.0%, outperforming state-of-the-art classifiers. To speed up the process and improve efficiency, GPU convolution operations were implemented.

9.2.5 Internet-of-Things

IOT has been defined by the radio frequency identification group as "the world-wide network of interconnected objects uniquely addressable based on standard communications protocols" (Gubbi et al. 2013). These objects, such as sensors can be embedded in various devices across diverse domains such as healthcare, environment and astrology and are continually collecting and communicating data. These data are often semi-structured and require processing and analysis to provide useful information (Riggins and Wamba 2015).

An example of IOT and big data analytics is urban planning and smart cities (Kitchin 2014). A smart city can consist of devices built into the urban environment such as utility, communication and transport systems. These devices can be used in real-time to monitor and regulate city flows and processes. The integration and analysis of the data produced from these devices could provide an improved understanding of the city that enhances efficiency and sustainability (Hancke et al. 2013) and further models and predicts urban processes for future urban development (Batty et al. 2012). Examples of such platforms to support the IOT within a smart city include ThingSpeak[5] which provides a cloud-based platform where sensor data can be uploaded and analyzed using MatLab and iOBridge,[6] which provides a hardware solution to connect to the cloud with developed Application Programming Interfaces (APIs) to allow integration with other web services. Multi-nationals such as HP and IBM are also investing in projects such as CeNSE[7] and Smarter Planet,[8] respectively. CeNSE is deploying a vast number of sensors used to track for a range of applications from monitoring use and location of hospital equipment to tracking traffic flow. It then gathers and transmits such data to computing engines for analysis in real-time.

[5]ThingSpeak. https://thingspeak.com/pages/learn_more. Accessed: 2017-02-03.

[6]ioBridge. http://connect.iobridge.com/. Accessed: 2017-02-03.

[7]CeNSE. http://www8.hp.com/us/en/hp-information/environment/cense.html#.WJCsHbaLR0K. Accessed: 2017-02-03.

[8]IBM Smarter Planet. http://www.ibm.com/smarterplanet/us/en/. Accessed: 2017-02-03.

9.2.6 *Finance*

Financial institutions are adopting a data-driven approach with the aim of improving their performance, service and, as seen with the financial crash in 2008, their risks (Fan et al. 2014). Financial data can be in a structured or semi-structured form; such data includes stock prices, derivative trades, transaction records and high-frequency trades (HFT). A study by Seddon and Currie (2017) proposed a model for applying big data analytics in HFT. HFT uses algorithmic software to perform trades built upon advanced technological infrastructure with a focus on speed to process and leverage vast amounts of financial data (Aldridge 2009). This study analyzed big data and its impact upon financial markets. An important discussion, applicable to all application areas is data security and privacy. With high volumes of data used in analysis, questions need to be addressed around data security protection, intellectual property protection, personal privacy protection, commercial secrets and financial information protection (Chen and Zhang 2014).

9.3 Computational Challenges

At the heart of many of the computationally intense applications lies pattern matching and machine learning:

- Machine learning
- Deep learning
- Pattern matching
- Image/video/audio processing
- Sentiment analysis
- Natural language processing

Recent advances in high-performance computing has encouraged the field of deep learning to move out from research laboratories and become a commercial opportunity. Deep learning, driven by research centers and initiatives such as the Google Brain project,[9] has projected to become a multi-billion pound industry by 2024 (Tractica 2015; PR Newswire 2016), finding potential enterprise applications in areas of finance, advertisement, automotive, medical and other end-user applications. An enabler for this projected growth is in research and development of infrastructures, software and hardware technologies optimized for deep learning solutions.

[9]Google Brain Team. https://research.google.com/teams/brain/. Accessed: 2017-02-03.

9.4 High-Performance Computing Solutions

A background into different approaches is provided in this section. It should be noted that different application domains will have varied computational demands (Singh and Reddy 2014). The sections below discuss high-performance computing solutions ranging in computational performance.

9.4.1 Graphics Processing Units (GPU) Computing

Graphics processing units as the name suggests, are custom devices consisting of many processing cores or co-processors that have been tailored for processing the vast computational and memory requirements for graphics rendering and image processing. They enable highly mathematical and computationally intense functions to be performed at an accelerated rate due to the parallel computational units at the heart of their structure. The ability to offload computation most suited to parallel operations, while maintaining a great level of flexibility and scalability is a leading benefit of GPU-based computing over sequential operation CPU-based computing (Blayney et al. 2015; Melanakos 2008; Fan et al. 2004). However, the scale of the benefits depends strongly on the nature of the computations.

The application and use of GPUs has gone far beyond computer graphics and gaming, although expansion these markets have certainly reduced the cost of GPUs, making them a more affordable and thus widespread technology (Fan et al. 2004). The terms General-Purpose computation on Graphics Processing Units (GPGPU) and GPU Computing have arisen which signifies that the processors have a broad range of potential applications.

NVIDIA, is a market leader GPU producer, providing a range of GPU processors, boards and platforms.[10] The power of their GPUs can be harnessed through NVIDIA's own Compute Unified Device Architecture (CUDA) parallel computing platform. This technology has been used in a range of applications spanning gaming, mobile, personal computers through to high-performance computing, and deep learning. For example, in bioinformatics there have been a large number of CUDA-based tools developed for accelerating sequence processing and analysis (Klus et al. 2012; Liu et al. 2012, 2013). Although GPU computing is a promising direction for bioinformatics, memory handling and slow data exchange between CPU and GPU processors can still cause challenges (Starostenkov 2013).

In the area of deep learning, NVIDIA sees a market extending its capabilities in the area of accelerating Artificial Intelligence (AI) algorithms (Azoff 2015)

[10]NVIDIA. http://www.nvidia.co.uk/page/home.html. Accessed: 2017-02-03.

in industries such as automotive, internet, healthcare, government, finance and others.[11] They are clearly positioning themselves for the expected growth in the big data market.

9.4.2 Field Programmable Gate Arrays

FPGAs are integrated circuits which enable a level of programmability. Their structure consists of an array of programmable logic blocks containing computational units, memory and interconnections that can be fully preconfigured. They sit between highly programmable digital signal processing chips and custom design ASICs, providing a balance of flexibility with parallel custom designed operations. They offer an experimentation and development platform to design and refine solutions. Yet they also provide enterprise solutions for applications in which a certain degree of reconfiguration may be required. However, unlike CPUs and GPUs this reconfiguration cannot be done totally on the fly and takes a level of reprogramming the device. Where there are advantages is when there is a large number of repetitive operations that are suited to parallel implementation, such examples are in image processing, pattern matching, or routing algorithms. In such cases FPGAs can be orders of magnitude faster compared to other platforms. The content below provides an overview of some examples of FPGAs in use.

FPGAs can offer possible solutions to computational challenges in bioinformatics and molecular biology (Ramdas and Egan 2005). A major computational challenge in genomics is in sequence alignment. The Smith–Waterman algorithm is a database search algorithm suited for protein sequence alignment. However, it is computationally intensive and the complexity increases quadratically as the dataset increases. Dydel and Bala (2004), present an implementation of it on FPGA. Tan et al. (2016) also present a FPGA-based co-processor to speed up short read mapping in HTS, reporting a throughput of 947 Gbp per a day, while providing better power efficiency.

Another aspect that can benefit from computational enhancement is in the image processing component in Genomic Microarrays. In these examples, sequencing is not being performed, however, genetic markers are being looked for that respond to known chemical interactions leading to a change in colour in the array, depending on the level of expression. Rodellar et al. (2007) present such a device, tailored to be portable so to make it applicable in regions remote from core healthcare provision. An implementation of the CAST algorithm used for detecting low-complexity regions in protein sequences is described by Papadopoulos et al. (2012). Significant speed-up in computations in the region of 100× where observed. These examples are

[11] NVIDIA: Artificial Intelligence and Deep Learning. http://www.nvidia.co.uk/object/deep-learning-uk.html. Accessed: 2017-02-03.

not in themselves related to big data, however, they have relevance in the context of personalized medicine in which such data can routinely form part of a heterogeneous patient dataset.

9.4.3 Cloud Computing Platforms

The National Institute for Standards and Technology (NIST) defines Cloud computing as "a pay-per-use model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction".

Foster et al. (2001) pioneered an idea of Grid computing which constitutes a large-scale distributed resource sharing under specified rules among the users and/or organizations. This idea was based on other known technologies of the time such as distributed computing. Grid computing proved to be useful in many scenarios, especially, for large-scale scientific computations (Di et al. 2012).

The concept of 'Clouds' as a similar but yet different way of distributed computing has been popularized by Amazon[12] in 2006. Armbrust et al. (2010) compare Cloud computing to other similar computing concepts in their work. Hence, they claim that although Grid computing offers protocols to share distributed resources, Cloud computing has advanced forward by offering "a software environment that grew beyond its community" (referring to the high-performance community).

Cloud computing has become a strong industry enabling a range of different services to be deployed typically by a pay-per-use cost model providing scalability in computing performance, storage and applications. Their expandability and sheer flexibility of services can provide a cost effective option for organizations in which the cost for development and maintenance for in-house solutions does not make business sense. Furthermore, cloud services can provide tools such as project and data management tools to aid in collaborations, provision of security and regulations in accessing shared data and analytical resources for the visualisation and understanding of datasets.

Cloud services fall under three different categories depending on the extent of the service provided:

- Infrastructure as a Service (IaaS) – Providing access to the core computing and storage infrastructure.
- Platform as a Service (PaaS) – Users can develop or build upon libraries and existing core platforms, and these solutions run on the cloud infrastructure.
- Software as a Service (SaaS) – Users access applications that form part of the cloud infrastructure.

[12] Amazon About AWS. https://aws.amazon.com/about-aws/. Accessed: 2017-02-03.

Some of the first adopters of big data in cloud computing are users that deployed NoSQL and Hadoop clusters in highly scalable and elastic computing environments provided by vendors, such as Google, Microsoft, and Amazon. An overview of the key market players is summarised as follows.

9.4.3.1 Amazon Web Services

Amazon Web Services (AWS) are the strongest competitors in cloud services (Leong et al. 2016), entering the market in 2006 and offering a range of relatively cost effective solutions. Their Amazon Elastic Compute Cloud (EC2) provides a scalable IaaS cloud service,[13] offering users a simplistic interface to their computing infrastructure. PaaS services are also supported. AWS have added Amazon EC2 Elastic GPUs to their provision allowing performance enhancements.

9.4.3.2 Microsoft Azure

Microsoft Azure provides both PaaS and more recently IaaS services.[14] The Azure platform offers functionality to integrate models, analyze data and visualization tools to scale data analysis. The Microsoft Azure model has been described in Gannon et al. (2014) as "layers of services for building large scale web-based applications". These layers communicate across various levels including the hardware level, utilizing data centers worldwide for computation and content delivery. The 'fabric controller' acts as the kernel of the Azure operating system. It performs tasks such as monitoring and managing the virtual machines and hardware resources that make up the Azure system.

9.4.4 Deep Learning Libraries

Machine learning and, in particular, deep learning have become of immediate interest for companies and researchers alike. Such technology is finding its way into a range of products from speech recognition, image processing, search optimization, through to any application where there is a need or interest to understand behavior, images, speech and sentiment analysis. TensorFlow™ and other such systems can be a great enabler to develop such features.[15]

TensorFlow™ is an open source machine learning infrastructure originating from Google as part of their Google Brain project started in 2011. It formed part of

[13] Amazon EC2. https://aws.amazon.com/ec2/. Accessed: 2017-02-03.

[14] Microsoft Azure. https://azure.microsoft.com/. Accessed: 2017-02-03.

[15] TensorFlow™. https://www.tensorflow.org/. Accessed: 2017-02-03.

Table 9.1 Deep learning libraries

Library	Detail
TensorFlow™	Open source machine learning infrastructure originating from Google as part of their Google Brain project started in 2011 (https://www.tensorflow.org/)
MXNet	Flexible library (http://mxnet.io/) which supports multiple languages (C++, Python, R, Scala, Julia, Matlab and Javascript), can operate on personal CPU/GPU setups through to distributed and cloud platforms (including AWS, Google Compute Engine (https://cloud.google.com/compute/), Microsoft Azure)
Caffe	A deep learning framework developed by the Berkeley Vision and Learning Center (http://caffe.berkeleyvision.org/) (Jia 2014). Offers a competitive high-performance convolutional neural networks solution
Theano	A Python-based library with a focus on enhancing mathematical computation of multi-dimensional arrays (http://deeplearning.net/software/theano/)
Torch	A scientific framework for machine learning (http://torch.ch/)

the Google's Machine Intelligence research organization with its focus on machine learning and in particular deep neural networks. A key feature of TensorFlow™ is its sheer scalability and flexibility. It facilitates distribution of computations over a range of devices and platforms, from mobile devices and desktops, through to large scale infrastructures consisting of hundreds of machines or thousands of GPU devices (Abadi et al. 2015). More recently it has been incorporated within AWS Elastic Cloud (Amazon EC2) provision. It is part of their Deep Learning Amazon Machine Image (AMI) and is just one of a suite of deep learning libraries included (see Table 9.1).

9.5 The Role for Custom Hardware

Do we need to look at big data at the micro level or at the macro level? For example, genetic sequencing, particularly as part of next generation sequencing, requires a substantial computational overhead in the alignment of the small reads coming from the initial sample analysis. From this alignment the DNA sequence of smaller exome components can then be used to determine conditions and states of disease. Opposite to this are huge datasets of genomic data across thousands of people ranging in phenotype and genomic marker such as exome sequences. Gathering such huge expanses of genetic data and combining this with other associated information offers huge opportunities in disease stratification, biomarker discovery and drug development (Raghupathi and Raghupathi 2014). This is clearly big data at the macro level. So the question lies – would the same high-performance computing suit both applications? This particular example is further complicated by the size of even a single DNA sequence. Uploading such a file-size to a cloud-based system in itself

presents challenges. Techniques have been developed to look at easing storage of such genetic information. One particular approach is with compression algorithms to find an efficient method to represent the data (Qiao et al. 2012). Such a method needs to be loss-less, fast, and effective.

Another consideration could be the need for secure solutions which keep data local, although cloud services such as AWS take great measures to keep their services secure. Establishing a custom system incurs a significant investment and maintenance overhead, and would be difficult to scale up. However, big data computations pose an ever increasing challenge in meeting performance needs. In particular, deep learning is an area of machine learning showing great commercial prospect. The next sections look at some of the deep learning solutions available.

9.5.1 Deep Learning

TensorFlow™ and other deep learning libraries (Table 9.1) combined with cloud services provide a platform to develop and create deep learning solutions, leading on to commercial opportunities. However, despite the great flexibility and scalability advantages of such a system, is there a possibility that a hardware-based solution might provide the better solution? This of course depends strongly on the application at hand and the limitations and challenges associated. Nevertheless, deep learning is a component of machine learning with great commercial interest. fpgaConvNet (Venieris et al. 2016) is a framework for mapping convolutional neural networks, a form of deep learning, onto FPGAs. The authors relate to the computational issues presented in convolutional networks, in particular, the classification computation overhead and the rapid scaling in complexity. CNNLab (Zhu et al. 2016), is another parallel framework for deep learning neural networks that distributes computation to both GPUs and FPGAs. Microsoft Azure has also incorporated FPGAs within their cloud platform (Feldman 2016). Woods and Alonso (2011) have developed an FPGA based framework for analytics on high-rate data streams. The next section looks further at enhancing cloud performance through incorporating custom hardware provision.

9.5.2 ASIC Enhanced Cloud Platforms

Nervana[16] has developed a platform for deep learning that is powered using a custom ASIC engine accessed through a cloud platform. They state that their cloud solution enables industry commercialized deep learning solutions. The platform they provide

[16]Nervana. https://www.nervanasys.com/intel-nervana/. Accessed: 2017-02-03.

is described by them as a full stack solution for "AI on demand", optimized at each level.

Nervana Neon is an open source Python-based scalable deep learning library. The Nervana Engine is custom ASIC hardware optimized for machine learning and in particular deep learning. They promote high-speed data access with high bandwidth memory, reaching speeds of 8 Terabits per second for memory access. Additionally, on-chip memory is large (32 GB) to meet the excessive storage requirements for machine learning. The core computational power is achieved through a sea of multipliers supported with local memory, without a reliance on cache memory. Nervana have paid great attention to data transfer across the chip including communication pipelines tailored for machine learning operations. One key aspect of this is the design allowing ASICs to be interconnected directly without reliance on Peripheral Component Interconnect Express (PCIE) buses which cause data flow bottlenecks. Nervana Engine is set to be released in 2017 and hopes to establish a place in the top deep learning technologies (Schneider 2017).

9.5.3 ASIC Deep Learning Processors

However, Nervana are not the only ones interested in this market with others are providing custom machine learning processing engines.

One of the most interesting areas in developing on-chip processing is based on the operation of the human brain, termed Neuromorphic chips. In this field, Spiking Neural Networks (SNN) are used to form the computations. The SpiNNaker Project is one example (Sugiarto et al. 2016) and forms part of the Human Brain Project.[17] The Darwin Neural Processing Unit is another exciting example of an ASIC co-processer based on SNN (Shen et al. 2016). Through the very nature of how SNN operate they may lend themselves more closely to machine learning and therefore show great promise in this area (Elton 2016).

9.6 Discussion

Big data and its analysis have the potential to provide insight into many diverse domains. The wealth of data collected at such a vast scale has led to the need for computationally intensive solutions to find useful information hidden in the chaos. The applications for such analysis are far reaching, from surveillance, finance, IOT, and smart cities through to personalized health. Potential of such applications include clinical decision support systems, personalized medicine for healthcare, distribution and logistics optimization for retail and supply chain planning for

[17] Human Brain Project. https://www.humanbrainproject.eu/en_GB. Accessed: 2017-02-03.

manufacturing (Sagiroglu and Sinanc 2013). However, even within each example, applications will have different needs in terms of data growth, infrastructure and governance along with integration, velocity, variety, compliance and data visualization (Intel 2012). A number of challenges still need to be addressed such as handling structured and unstructured data in real/near-time at a volume whereby traditional data storage and analysis approaches are not applicable (Zikopoulos and Eaton 2012). Furthermore, as big data analytics becomes mainstream, important issues such as data governance, guaranteeing privacy, safeguarding security, increased network bottlenecks, training of skilled data science professionals, development of compression technologies and establishing standards will require urgent attention (Intel 2012).

Big data analytics and applications are still in the early stages, however, the continuation of technology and platform improvement such as Hadoop, Spark, NoSQL coupled with the development of new analytical algorithms and infrastructure will contribute towards the maturing of the field. Companies such as Nervana are developing custom hardware to work in tandem with their cloud platform to accelerate deep learning. This is one field in which hardware developers can create impact for cloud computing infrastructure and big data analytics. Recently, Microsoft (Feldman 2016) announced the inclusion of Altera FPGAs within their Azure cloud service with the promise of creating an AI supercomputer. Microsoft does not currently plan to use the FPGAs for training neural networks, using GPUs instead for offline training. At present, they see FPGAs providing effective acceleration for evaluating already trained neural networks.

Qualcomm, recognize that their consumers require on-device solutions that do not rely fully on cloud services. Their machine learning platform is implemented on their Snapdragon Neural Processing Engine. The example here highlights that data analytics is a challenge that may not always be resolved through scalable cloud services, but as applications require more computationally intensive data analytics, some of this workload may need to be shared between on-device and cloud-based services. Other companies are also active in this area (Table 9.2) and seemingly there is a strong market for this level of on-device processing. Furthermore, there have been exciting advances happening in the area of Neuromorphic chips for machine learning. It will be interesting to see how this technology impacts the deep learning market.

Clearly, each computational solution offers unique opportunities for overcoming the challenges of big data. FPGA and ASIC solutions can provide computational benefits under certain conditions and as demonstrated through companies such as Microsoft and Nervana they can form a key part of a high-performance cloud platform. Conversely, they play an important role for on-device big data analytics with companies such as Qualcomm and Intel investing largely in developing the next generation of AI chips. In each example the solutions have been tailored for the ever growing market of big data and deep learning. Meeting these challenges will have great impact to applications in the future, advances in healthcare, smart cities, security, automotive industry among other examples forming part of our daily lives.

Table 9.2 Deep learning ASIC processors

Product	Detail
Qualcomm Snapdragon Neural Processing Engine	Deep learning toolkit for mobile and edge devices from Qualcomm Technologies (https://www.qualcomm.com/invention/cognitive-technologies/machine-learning)
Qualcomm Zeroth SDK	On-device machine learning platform (Vicent 2016)
Google's Tensor Processing Unit	Part of Google's drive for deep learning solutions (Osborne 2016). Accelerator ASIC developed to be accompanied by their TensorFlow™ library
Intel Xeon Phi product family – Knights Mill/Knight Landing/Knights Crest	Family of high-performance custom ASICs for machine learning (Hruska 2016). Their product development includes bringing together Nervana's chip technology (Intel acquired Nervana in 2016) together with Xeon processors to produce their Knights Crest chip

References

23andMe (2015) DNA genetic testing & analysis. 23andme. Available via https://www.23andme.com/. Accessed 06 Feb 2017

Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, (30 additional authors not shown) (2015) Tensorflow: large-scale machine learning on heterogeneous distributed systems. arXiv:1603.04467 [cs.DC]. Available via https://arxiv.org/abs/1603.04467

Aldridge I (2009) High-frequency trading: a practical guide to algorithmic strategies and trading systems, 2nd edn. Wiley, Somerset

Almalki M, Gray K, Sanchez FM (2013) The use of self-quantification systems for personal health information: big data management. Health Inf Sci Syst 3(Suppl 1):1–11

AncestryDNA™ (2016) DNA tests for ethnicity & genealogical DNA testing. AncestryDNA™. Available via https://www.ancestry.co.uk/. Accessed 06 Feb 2017

Armbrust M, Fox A, Griffith R, Joseph AD, Katz R, Konwinski A, Lee G, Patterson D, Rabkin A, Stoica I, Zaharia M (2010) A view of cloud computing. Commun ACM 53(4):50

Azoff M (2015) Machine learning in business use cases: artificial intelligence solutions that can be applied. NVIDIA. Available via http://www.nvidia.com/. Accessed 06 Feb 2017

Baker M (2010) Next-generation sequencing: adjusting to data overload. Nat Methods 7:495–499

Batty M, Axhausen KW, Fosca G, Pozdnoukhov A, Bazzani A, Wachowicz M, Ouzounis GK, Portugali J (2012) Smart cities of the future. European Phys J Spec Top 214(1):481–518

Blayney J, Haberland V, Lightbody G, Browne F (2015) Biomarker discovery, high performance and cloud computing: a comprehensive review. In: Proceedings of 2015 IEEE International conference on bioinformatics and biomedicine (BIBM), pp 1514–1519

British Gas (2017) How data can personalise your energy. British gas. Available via https://www.britishgas.co.uk/. Accessed 06 Feb 2017

Chen CLP, Zhang CY (2014) Data-intensive applications, challenges, techniques and technologies: a survey on big data. Infor Sci 275:314–347

Chen M, Mao S, Liu Y (2014) Big data: a survey. Mob Netw Appl 19(2):171–209

Di S, Kondo D, Cirne W (2012) Characterization and comparison of cloud versus grid workloads. In: Proceedings of 2012 IEEE International conference on cluster computing (CLUSTER'12), pp 230–238

Dydel S, Bała P (2004) Large scale protein sequence alignment using FPGA reprogrammable logic devices. In: Proceedings of 14th International conference field programmable logic and application (FPL'04), pp 23–32

Elton D (2016) Neuromorphic chips: a path towards human-level AI. Singularity. Available via https://www.singularityweblog.com/. Accessed 06 Feb 2017
Erlich Y (2015) A vision for ubiquitous sequencing. Genome Res 25(10):1411–1416
Fan Z, Qiu F, Kaufman A, Yoakum-Stover S (2004) GPU cluster for high performance computing. In: Proceedings of 2004 ACM/IEEE conference on supercomputing (SC'04), pp 47–47
Fan J, Han F, Liu H (2014) Challenges of big data analysis. Natl Sci Rev 1(2):293–314
Feldman M (2016) Microsoft goes all in for FPGAs to build out AI cloud. TOP500 supercomputer sites. Available via https://www.top500.org/. Accessed 06 Feb 2017
Foster I, Kesselman C, Tuecke S (2001) The anatomy of the grid: enabling scalable virtual organizations. Int J High Perform Comput Appl 15(3):200–222
Gannon D, Fay D, Green D, Takeda K, Yi W (2014) Science in the cloud: lessons from three years of research projects on Microsoft Azure. In: Proceedings of 5th ACM workshop on scientific cloud computing (ScienceCloud'14), pp 1–8
Gubbi J, Buyya R, Marusic S, Palaniswami M (2013) Internet of things IoT: a vision, architectural elements, and future directions. Future Gener Comput Syst 29(7):1645–1660
Hancke GP, de Carvalho e Silva B, Hancke GP Jr (2013) Sensors. Role Adv Sens Smart Cities 13(1):393–425
Hashem IAT, Yaqoob I, Anuar NB, Mokhtar S, Gani A, Ullah KS (2015) The Rise of "big data" on cloud computing: review and open research issues. Inform Syst 47:98–115
Hruska J (2016) Intel announces major AI push with upcoming Knights Mill Xeon Phi, custom silicon. Extreme Tech. Available via https://www.extremetech.com/. Accessed 06 Feb 2017
Intel (2012) Big data analytics Intel's IT manager survey on how organizations are using big data. Intel. Available via http://www.intel.me/. Accessed 06 Feb 2017
Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T (2014) 'Caffe'. In: Proceedings of 22nd ACM International conference on multimedia (MM'14), pp 675–678
Kitchin R (2014) The real-time city? Big data and smart urbanism. GeoJournal 79(1):1–14
Klus P, Lam S, Lyberg D, Cheung MS, Pullan G, McFarlane I, Yeo GS, Lam BY (2012) BarraCUDA – a fast short read sequence aligner using graphics processing units. BMC Res Notes 5(1):27
Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. In: Proceedings of advances in neural information processing systems 25 (NIPS'12), pp 1097–1105
Kühnemund M, Wei Q, Darai E, Wang Y, Hernández-Neuta I, Yang Z, Tseng D, Ahlford A, Mathot L, Sjöblom T, Ozcan A, Nilsson M (2017) Targeted DNA sequencing and in situ mutation analysis using mobile phone microscopy. Nat Commun 8:13913
Leong L, Petri G, Gill B, Dorosh M (2016) Magic quadrant for cloud infrastructure as a service, worldwide. Gartner. Available via https://www.gartner.com. Accessed 06 Feb 2017
Li G, Bankhead P, Dunne PD, O'Reilly PG, James JA, Salto-Tellez M, Hamilton PW, McArt D (2016) Embracing an integromic approach to tissue biomarker research in cancer: perspectives and lessons learned. Briefings Bioinf bbw044, pp 1–13. Available via https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbw044
Lightbody G, Browne F, Zheng H, Haberland V, Blayney J (2016) The role of high performance, grid and cloud computing in high-throughput sequencing. In: Proceedings of 2016 IEEE International conference on bioinformatics and biomedicine (BIBM), pp 890–895
Liu Y, Schmidt B, Maskell DL (2012) Cushaw: a CUDA compatible short read aligner to large genomes based on the Burrows-Wheeler transform. Bioinformatics 28(14):1830–1837
Liu Y, Wirawan A, Schmidt B (2013) CUDASW++ 3.0: accelerating Smith-Waterman protein database search by coupling CPU and GPU SIMD instructions. BMC Bioinf 14(1):117
Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ (2012) Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol 30(5):434–439
Melanakos J (2008) Parallel computing on a personal computer. Biomedical Computation Review. Available via http://www.biomedicalcomputationreview.org/. Accessed 06 Feb 2017

Metzker ML (2009) Sequencing technologies – the next generation. Nat Rev Genet 11(1):31–46
Miller NA, Farrow EG, Gibson M, Willig LK, Twist G (16 additional authors not shown) (2015) A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases. Genome Med 7(1):100
Naccache SN, Federman S, Veeeraraghavan N, Zaharia M, Lee D (21 additional authors not shown) (2014) A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. Genome Res 24(7):1180–1192
Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R, Muharemagic E (2015) Deep learning applications and challenges in big data analytics. J Big Data 2(1):1
O'Driscoll A, Daugelaite J, Sleator RD (2013) 'Big data', Hadoop and cloud computing in genomics. J Biomed Infor 46(5):774–781
O'Dwyer E, De Tommasi L, Kouramas K, Cychowski M, Lightbody G (2016) Modelling and disturbance estimation for model predictive control in building heating systems. Energy Build 130:532–545
Osborne J (2016) Google's tensor processing unit explained: this is what the future of computing looks like. TechRadar. Available via http://www.techradar.com/. Accessed 06 Feb 2017
Papadopoulos A, Kirmitzoglou I, Promponas VJ, Theocharides T (2012) FPGA-based hardware acceleration for local complexity analysis of massive genomic data. Integr VLSI J 46(3):230–239
PR Newswire (2016) $1.77 billion deep learning market 2016 – global forecasts to 2022: Google is among the market. PR Newswire. Available via http://www.prnewswire.com/. Accessed 06 Feb 2017
Qiao D, Yip WK, Lange C (2012) Handling the data management needs of high-throughput sequencing data: speedgene, a compression algorithm for the efficient storage of genetic data. BMC Bioinf 13(1):100
Raghupathi W, Raghupathi V (2014) Big data analytics in healthcare: promise and potential. Health Infor Sci Syst 2(3):1–10. Available via https://link.springer.com/article/10.1186/2047-2501-2-3/fulltext.html
Ramdas T, Egan G (2005) A survey of FPGAs for acceleration of high performance computing and their application to computational molecular biology. In: Proceedings of TENCON 2005 – 2005 IEEE region 10 conference, pp 1–6
Riggins FJ, Wamba SF (2015) Research directions on the adoption, usage, and impact of the Internet of Things through the use of big data analytics. In: Proceedings of 2015 48th Hawaii International conference on system sciences, pp 1531–1540
Rodellar V, Díaz F, Belean B, Malutan R, Stetter B, Gomez P, Martínez-Olalla R, García-Rico E, Pelaez J (2007) Genomic microarray processing on a FPGA for portable remote applications. In: Proceedings of 2007 3rd Southern conference on programmable logic (SPL'07), pp 13–17
Sagiroglu S, Sinanc D (2013) Big data: a review. In: Proceedings of 2013 International conference on collaboration technologies and systems (CTS), pp 42–47
Schneider D (2017) Deeper and cheaper machine learning [Top Tech 2017]. IEEE Spectr 54(1):42–43
Seddon JJJM, Currie WL (2017) A model for unpacking big data analytics in high-frequency trading. J Bus Res 70:300–307
Shen J, Ma D, Gu Z, Zhang M, Zhu X, Xu X, Xu Q, Shen Y, Pan G (2016) Darwin: a neuromorphic hardware co-processor based on spiking neural networks. Sci China Infor Sci 59(2):1–5
Singh D, Reddy CK (2014) A survey on platforms for big data analytics. J Big Data 2(1):8
Starostenkov V (2013) Hadoop + GPU: boost performance of your big data project by 50x-200x? Network World. Available via http://www.networkworld.com/. Accessed 06 Feb 2017
Sugiarto I, Liu G, Davidson S, Plana LA, Furber SB (2016) High performance computing on SpiNNaker neuromorphic platform: a case study for energy efficient image processing. In: Proceedings of 2016 IEEE 35th International performance computing and communications conference (IPCCC), pp 1–8
Swan M (2013) The quantified self: fundamental disruption in big data science and biological discovery. Big Data 1(2):85–99

Tan G, Zhang C, Tang W, Zhang P, Sun N (2016) Accelerating irregular computation in massive short reads mapping on FPGA co-processor. IEEE Trans Parallel Distrib Syst 27(5):1253–1264

Terrell J (2015) Test-driving the brain could reveal early signs of Alzheimer's. The Conversation. Available via http://theconversation.com/. Accessed 06 Feb 2017

Townsend AM (2014) Smart Cities – Big Data, Civic Hackers, and the Quest for a New Utopia. W. W. Norton & Company, Edition Reprint

Tractica (2015) Deep learning software market to surpass $10 billion by 2024. Tractica. Available via https://www.tractica.com/. Accessed 06 Feb 2017

Van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C (2014) Ten years of next-generation sequencing technology. Trends Genet 30(9):418–426

Venieris SI, Bouganis CS (2016) fpgaConvNet: a framework for mapping convolutional neural networks on FPGAs. In: Proceedings of 2016 IEEE 24th annual International symposium on field-programmable custom computing machines (FCCM), pp 40–47

Vicent J (2016) Qualcomm's deep learning SDK will mean more AI on your smartphone. The Verge. Available via http://www.theverge.com/. Accessed 06 Feb 2017

Wang Y, McCleary D, Wang CW, Kelly P, James J, Fennell DA, Hamilton PW (2010) Ultra-fast processing of gigapixel tissue microarray images using high performance computing. Anal Cell Pathol (Amsterdam) 33(5):271–285

Woods L, Alonso G (2011) Fast data analytics with FPGAs. In: Proceedings of 2011 IEEE 27th International conference on data engineering workshops, pp 296–299

Xu Z, Liua Y, Meia L, Hua C, Chen L (2015) Semantic based representing and organizing surveillance big data using video structural description technology. J Syst Softw 102:217–225

Xu Z, Mei L, Hu C, Liu Y (2016) The big data analytics and applications of the surveillance system using video structured description technology. Clust Comput 19(3):1283–1292

Zhu M, Liu L, Wang C, Xie Y (2016) CNNLab: a novel parallel framework for neural networks using GPU and FPGA – a practical study with trade-off analysis. arXiv:1606.06234 [cs.LG]. Available via https://arxiv.org/abs/1606.06234. Accessed 06 Feb 2017

Zikopoulos PC, Eaton C (2012) Understanding big data: analytics for enterprise class Hadoop and streaming data. McGraw-Hill Osborne Media, New York