Big Data Framework

Firat Tekiner and John A. Keane
School of Computer Science,
The University of Manchester,
Manchester, UK
ftekiner@yahoo.com, john.keane@manchester.ac.uk

Abstract— We are constantly being told that we live in the Information Era – the Age of BIG data. It is clearly apparent that organizations need to employ data-driven decision making to gain competitive advantage. Processing, integrating and interacting with more data should make it better data, providing both more panoramic and more granular views to aid strategic decision making. This is made possible via Big Data exploiting affordable and usable Computational and Storage Resources. Many offerings are based on the Map-Reduce and Hadoop paradigms and most focus solely on the analytical side. Nonetheless, in many respects it remains unclear what Big Data actually is; current offerings appear as isolated silos that are difficult to integrate and/or make it difficult to better utilize existing data and systems. Paper addresses this lacunae by characterising the facets of Big Data and proposing a framework in which Big Data applications can be developed. The framework consists of three Stages and seven Layers to divide Big Data application into modular blocks. The aim is to enable organizations to better manage and architect a very large Big Data application to gain competitive advantage by allowing management to have a better handle on data processing.

Keywords: Big Data, data scientist, analytics, business intelligence, information management, strategy, hadoop

I. INTRODUCTION

We live in the Information Era – the Age of BIG data

As an example, Big Data’s significance and power became apparent when the results of the 2012 US Presidential Elections were announced. Complex analytics processing large data not only predicted the exact election results but may also have influenced it [3][4]. Further, leading business magazines and economical newspapers run frequent articles about Big Data’s success [5][6].

However, it should be recognised that Big Data is not something new, it has long been the playground of the elite. The aim was to maximise expensive CPU utilisation. As a result, it had a limited audience as computation and storage was expensive and difficult to utilise requiring detailed systems knowledge where capacity doubles every 18 months [7].

In the Big Data era, computation and storage is cheap per TB. Therefore, with ever-growing computational capabilities, system utilisation is no longer as critical a factor. It is now feasible to use more computational power to do the same work (hence with lower utilisation). At the same time, the amount of data that needs processing has been increasing exponentially in the past decade as a result of improvements in data generation and storage capacity [1]. Above all, programming tools and methodologies have matured with globalisation and the Internet. It is increasingly feasible to reuse code (and also share it), therefore, the focus has moved to integrating codes created by different communities.

High performance network capacity, that provides the backbone for high end computing systems, has not increased at the same rate as processing and storage capabilities. Therefore, the constraint in computation has simply shifted from moving data to a big supercomputer, to moving the application to many smaller computers where the data resides (function shipping rather than data shipping). Programming such an approach is not new, the application is executed where the data is kept in a loosely coupled and highly distributed architecture [8].

In contrast, Relational Database Management Systems (RDBMS) tend to provide access to data as one Big Data silo based on efficient closely coupled systems. Structured Query Language (SQL) is the de-facto method to access databases as it provides relatively easy access to data at different levels within organisations. It is common to see low-level programmers and high level business analysts sharing the same piece of SQL and understanding, or trying to understand it.

This sharing model has its limitations and cannot exploit and handle the massive increase in static non-changing data. Recently, there has been an increase in NoSQL approaches to overcome these weaknesses [9]. Despite their relatively recent emergence, there are now more than one hundred NoSQL approaches that specialize in management of different multi-modal data types (from structured to non-structured) and with the aim to solve very specific challenges. Most are powered by the Map-Reduce paradigm that came from Google, which is based on a massively distributed architecture that exploits cheap commodity hardware. As a result, the need for efficient mechanism for storing and processing data is eliminated. It is in fact cheaper to duplicate (for reliability) and to over-compute (process duplicate data) as communication is relatively more expensive than storage and computational resources (and this gap is increasing).

1 Firat Tekiner has an honorary research fellowship at the School of Computer Science, University of Manchester.  
2 We bemusedly note that “big” has to an extent replaced “very large” from a previous generation – both remain undefined in any quantitative senses and seem to mean “whatever data amount challenges the state-of-the-art”.

2013 IEEE International Conference on Systems, Man, and Cybernetics
Despite falling hardware and system development costs, the size and complexity of the systems have been increasing to process more data in shorter time periods resulting in higher staff and management overhead; consequently, the total cost of ownership (TCO) of systems remains approximately the same [10].

Big Data is perceived as “the new driver of competitive advantage” [11]. Big Data applications have high volume, high variety and high velocity as findings are expected to be delivered very quickly [12]. In reality, Big Data is largely driven by the need to analyze massive volumes of data to gain competitive advantage and to use previously intractable processes to find information/relationships [13] [14]. This also provides more opportunities to bring in applications and fields to enrich data further to provide even better analysis. However, the wide technology gap between industrial applications and decision makers remains a challenge. Decision makers need to understand the data and technologies better in order to extract information to aid strategic decision making. Therefore, we conjecture that firms that bridge this gap quickly should gain significant competitive advantage by unearthing information that is not (immediately) available to others.

Nonetheless, despite its popularity, both the characteristics of “Big Data” applications and how they fit in a larger business context remain unclear [15]. For example, Big Data means different things to a DBMS vendor and a hardware vendor. They focus on solving problems solely in their area rather than the whole picture. As a result, there has been criticism that Big Data is simply hype. We disagree and believe Big Data offers a paradigm shift in both technical challenge and business opportunity – and requires this hybrid view.

Isolated applications processing data in a standalone fashion are insufficient; an approach is needed to make it easier to combine knowledge with Big Data analysis to provide an integrated enterprise level solution. This can be achieved only by marrying qualitative and quantitative analysis. Furthermore, Data Sources together with the underlying systems needs to be taken into account whilst integrating this information to achieve the velocity and volume aspects of the data. Therefore, there is a need to bridge the gap between domain experts, statisticians and computational methodologies [16].

The purpose of this paper is to move towards a more comprehensive definition of what “Big Data” really is by identifying and reviewing its characteristics; derived from this we propose a framework that encompasses all layers of a Big Data application and by defining lines of Big Data with a strategic view. The framework therefore, aims to synthesise and analyse all relevant aspects from users to Systems that define a Big Data environment. This work attempts to suggest future directions that such systems will take and how organizations can tractably and reliably increase their competitive advantage by exploiting advances in Big Data analysis.

The structure of the paper is as follows: in Section II related background is considered; the Big Data framework is proposed in Section III; finally, Section IV presents conclusions.

II. BACKGROUND

There have been many attempts to provide a standard programming environment, such as Message Passing Interface (MPI) and OpenMP, but by definition these do not solely target performance exploitation. Therefore, the paper will use the Map-Reduce paradigm as the basis for discussion as it has acted as the catalyst for increasing Big Data uptake despite it not being the sole model for Big Data.

The combination of cheap storage/computation complemented with the Map-Reduce programming methodology has enabled non-traditional High Performance Computing (HPC) users to process massive amounts of data in a timely manner. Furthermore, availability of open-source java applications and Map-Reduce resulted in very rapid uptake and use. Google’s success with Map-Reduce, followed by Yahoo’s Pig and Facebook’s Cassandra applications, has attracted much research and industry attention [17].

A. MapReduce and Hadoop

Map-Reduce was introduced by Google in order to process and store large datasets on commodity hardware. It provides a programming paradigm which allows useable and manageable distribution of many computationally intensive tasks. As a result, many programming languages now have Map-Reduce implementations which extend its uptake. On the other hand, Hadoop is a highly popular open-source Map-Reduce implementation by the Apache Foundation [17]. With the popularity of the hadoop applications there have been many complementing applications developed by the open source community and packaged up under apache foundation [18].

Map-Reduce involves two main parts, a map operation, where a simple function is used to emit key/value pairs in parallel similar to using primary keys in the relational database world. Once the data to be processed is mapped into key/value groups then a reduce operation is used to apply the core processing logic to produce results in a timely manner [19]. The simple concept of Map-Reduce removes many traditional
challenges in HPC to achieve fault tolerance and availability. Therefore, it paves the way for development of highly parallel, highly reliable and distributed applications on large datasets.

III. BIG DATA FRAMEWORK

A. Big Data Characteristics

Big Data is not only driven by the exponential growth of data but also by changing user behaviour and globalization. Much more time is being spent online and using mobile devices. Furthermore, globalization of the marketplace increases competition. As a result, organizations constantly look for opportunities to increase their competitive advantage in an increasingly competitive market place by using better analytical models. Hence, it is necessary to present findings in a more clear and concise form. In turn, there has been a commensurate increase in business intelligence applications that allow better reporting and visualization of the data.

To derive the framework, we firstly define the characteristics of “Big Data”:

1. Data/Processing Volume and Scale
2. Variety and Heterogeneity of Data/Sources
3. Speed and Timeliness of Information Requirement
4. Targeted Services, Products, Solutions and Applications
5. Data Presentation, Usability and Interpretation
6. Data Privacy, Error Handling and Security

Data volume has been increasing exponentially: up to 2.5 Exabytes of data is already generated and stored every day. This is expected to double within 40 months by 2015 [20]. As always there remains the challenge to process such volume.

The variety and heterogeneity of data sources and storage has increased, fuelled by the use of cloud, web and online computing. The challenge then becomes identification of data that will add value, and hence increase information content and competitive advantage [21]. Clearly, currency of information is crucial as analytics derived from new data is usually more valuable than old.

For example, consider an online business which analyses every click on its website. An advert or offering is made based on user’s movement and activities whilst browsing and shopping online. However, the adverts can be better targeted and the customers better segmented if customer profiles can be updated and integrated in sub-second intervals [22]. There are different patterns to the data and it is presented in different shapes. For example, management requires reporting and statistical analytics to be made available based on new data in order to be able to respond rapidly to changing requirements [21]. As these analytics provide predictive insight, the resulting decisions are both more robust and timely.

However, a shortage of skills and immature tools makes it a daunting task for organizations to present and interpret this newly discovered information and capability. Current hierarchical management models introduce difficulties in dynamic development and adaptation in an ever changing market place. Therefore, organizations that can respond and employ talents to understand, analyze, process and manage this information life cycle will lead the way. This is what has generated the interest, and hype, associated with big data. As a caveat, while integration of data provides many advantages, a significant associated risk is data privacy and ownership of the data. This is usually omitted or not understood at this stage and sometimes later) as the rush is to gain competitive advantage.

The paper argues that a Big Data application is the orchestration of all the software and hardware systems within the enterprise that generates and processes data. It means something different for each person, application or organization. For example, even a click whilst browsing a web page can be an input or a heartbeat packet that is sent over the network to inform that a system is still up and running. Up to now it was infeasible to store or process at this level of information. However, the information era is changing this.

B. Framework

There are formal approaches to project management that provide a methodology to manage Information Systems and drive strategy in Organizations; to name a few, Open Group’s TOGAF, IBM’s Zachman and Gartner’s methodology [23]. However, these are not designed to provide a framework that has a data focus, rather their aim is to provide methodologies to manage large information systems. In addition, [24] proposes a framework that looks into Big Data governance with an aim of managing people and policies. In contrast, having identified the characteristics of Big Data, this paper aims to define a framework that captures all the stages of a Big Data application with a strategic point of view of focusing on data. Although, [25] provides a Big Data methodology with a data focus it does not take into consideration the systems aspect of a Big Data environment. Furthermore, there is a need to bridge strategic decision making and real life scenarios. This paper aims to fill this space.

Without a coordination and structuring framework there is likely to be much overlap amongst applications, duplication in stored information and confusion around the responsibilities of each business unit and application. The framework here seeks to document the borders of each modular block to allow gaps to be spotted in a Big Data application and provide solutions by closer integration. Further, it aims to highlight how Map-Reduce can be included into the different stages and layers of the Big Data application life cycle. Whilst doing this, all surrounding issues and approaches are considered.

The framework should ultimately provide a basis to develop and manage Big Data applications whilst identifying strategies based on core competencies and weaknesses. In addition to the 7 layers identified in Figure 2, the process as a whole can be summarized in 3 main stages as below:

- STAGE 1 ➔ Multiple Data Sources - Choose the Right Data [26]
- STAGE 2 ➔ Data Analysis and Modelling
- STAGE 3 ➔ Data Organization and Interpretation
Stage 1 is concerned with acquisition and filtering of data by applying correct metadata and processes. Multiple data sources are integrated and transformed to add meaning to the data. This process is the major source of added-value (to data) and allows organizations to gain competitive advantage.

Stage 2 then uses the information prepared in Stage 1 to apply analytics and predictive models to find relationships and patterns that were not initially known. The level of intelligence applied depends on the computational capabilities and skill-set available together with the business requirements. Big Data uses internal and external datasets from a variety of sources to provide information to aid strategic decision making to gain competitive advantage. It allows focus on the current and the future rather than traditional historical reality. Whilst doing this, it further requires cross-functional collaboration at both business and technical level (data sources and systems) [27].

Stage 3 then deals with modelling the source information and mapping the data to the target model whilst interpreting the meaning of the newly discovered information. The relational data model does not naturally accommodate the unstructured and heterogeneous data sources that are expected to be available in Big Data applications [28]. Therefore, there may not be an upfront model whilst organizing the source data to the target. As a result of this, there have been a large number of applications that focus on providing access to these data sources via NoSQL without using SQL. They attempt to create indexing schemes similar to RDMBS and provide quick access to data residing in the Hadoop file system [29][30].

Presentation and visualisation of data is an important task. The NoSQL option changes the dynamics in terms of accessing and presenting the data. With increasing data to be analysed and processed, therefore, output needs to address both clarity and precision of presentation. In addition, interpretation of results is a major challenge that requires highly skilled staff.

The Processing stages described map onto the 7 layers of the framework. Each application may focus on different layers and may not employ all parts of it. A Big Data application then becomes a major orchestrating effort whereby a large number of moving parts needs to be composed to work seamlessly to achieve results that enable competitive advantage.

Figure 2. Big Data Framework
From a detailed perspective, how the Map-Reduce tasks will be applied depends on the application. As each map and reduce process can run in parallel, both can be used to speed up processing. Furthermore, at any given time, a number of Big Data applications can run at different layers or at different stages.

An important challenge is to bring together and map the relational database model with columnar, key-value stores and unstructured data. For example, Banks are experts about their customers; such information may be multiplied in value if joined together with unstructured sources [5]. In addition, Business Intelligence and Reporting applications requiring aggregations on a certain field are best served by DBMS that employ a columnar storage. Given that Business Intelligence traditionally uses RDBMS accessed via tools based on SQL, a change is needed.

Whilst the framework looks at and across all dimensions of the problem, almost all current Big Data approaches are silo-based without coherent linkage or integration. The modelling and mapping layer aims to do this with respect to data. For example, how the system and storage resources could be shared amongst different applications that would come under the Big Data framework is not a primary aspect of the design. This resource scheduling and maintenance needs to be managed at the system layer. In terms of storage, the challenge is bigger due to recent improvements in the medium; hybrid solid state, optical and hybrid disks operate at various speeds in addition to slower archiving systems. Separately, the data layer is expected to manage different data sources, handle data lineage and eliminate duplication that would otherwise be inevitable in an island of applications. This gap will grow further as storage struggles to cope with data growth and the Hadoop File System (HDFS) provides a cheaper yet reliable and performant alternative to current storage systems [31]. In addition, data privacy and security [32] aspects could be more easily managed within this layer under one common enterprise wide policy.

It has to be noted that, there are components which cannot be divided, such as SQL and DBMS or DBMS and hardware. For instance, modern DBMS make explicit use of and manage the underlying hardware. As a result of this there will also be overlaps and islands in an organisation. Hence, the application of the framework and abstraction based on the different layers.

IV. CONCLUDING REMARKS

Processing larger datasets has become increasingly possible over the past few years for a much larger community, not least via the development of the Map-Reduce paradigm. Map-Reduce enable the power of parallel computing to be available to standard data analysis tasks⁴.

As mentioned before, the main challenge in applying many islands of Big Data applications is to identify the defining lines of each application and their inter-relationship, in order to manage and integrate this spread within an organisation. To clarify what Big Data really is, it is the enterprise data processing environment for heterogeneous data and computational sources in a timely manner to gain competitive advantage. This results in the processing of high volume of data and presenting this in a concise and clear manner to aid operational and strategic decision making.

Due to the immaturity of the field, there is little or no coordination across Big Data silos (applications). Big Data does not only require in-depth systems and data expertise but also requires strategic insight due to the nature of the applications [11]. Such applications are evolving very quickly and designed to aid strategic decision making by responding rapidly to changes in market place. Organisations that are very hierarchical and bureaucratic may initially struggle to compete with the new economy companies [33]. This is evident from the fact that the companies that successfully apply Big Data applications are the likes of Google, Amazon, Yahoo and Facebook who are the leading new economy companies.

There is a lack of the multi-faceted role skills (analyst, developer, architect and management) required to orchestrate such applications [16]. The framework proposed here aims to document and structure this gap and provide a starting point for practitioners, analysts and management to develop and exploit their Big Data applications.

The framework can be seen as a cube corresponding to the levels. Each face represents an important level within the Big Data space, while the cube as a whole represents the entire Big Data space and the integrated whole we believe to be essential for effective deployment and evolution of the associated applications. To achieve this requires efficient and effective orchestration, integration and coordination of skills that address the challenges both within and across all seven levels defined in the previous section. This then needs to be further complemented by novel management and decision making strategies [34][35]. Thus there is a need for more technical managers and decision makers, and the lack of people with analytic skills is likely to be a challenge.

Big Data intersects with numerous domains including data integration, hardware and software, databases, Business Intelligence, system integrators and consulting firms [36]. The associated skill set is vast, and this is one of the challenges and confusions surrounding Big Data applications [15]. The associated scale is daunting and well indicates the need for integration to achieve effective and efficient use of Big Data [37]. Organizations need to be singularly focused whilst providing and employing Big Data solutions as management of all elements of the framework is challenging. Furthermore, there are many, and growing, numbers of applications that aim to use this improved “knowledge”. When all these aspects are considered, the combined issue is technical management and people management. Traditional management does not understand and cannot be expected to understand what can be achieved and what the related challenges are. Hence, the framework is proposed to aid decision making process and bridge the gap with business needs and technical realities.

---

⁴ It should be noted that Map-Reduce has weaknesses in that it is not, by design, general-purpose, but rather was designed for something very specific: keyword processing and access.