Chapter 21
The Heterogeneity Paradigm in Big Data Architectures

Todor Ivanov
Goethe University Frankfurt, Germany

Sead Izberovic
Goethe University Frankfurt, Germany

Nikolaos Korfiatis
University of East Anglia, UK

ABSTRACT

This chapter introduces the concept of heterogeneity as a perspective in the architecture of big data systems targeted to both vertical and generic workloads and discusses how this can be linked with the existing Hadoop ecosystem (as of 2015). The case of the cost factor of a big data solution and its characteristics can influence its architectural patterns and capabilities and as such an extended model based on the 3V paradigm is introduced (Extended 3V). This is examined on a hierarchical set of four layers (Hardware, Management, Platform and Application). A list of components is provided on each layer as well as a classification of their role in a big data solution.

INTRODUCTION

Undoubtedly the exponential growth of data and its use in supporting business decisions has challenged the processing and storage capabilities of modern information systems especially in the past decade. The ability to handle and manage large volumes of data has gradually turned to a strategic one (Chintagunta et al., 2013). Meanwhile, the term "Big Data" (Diebold, 2012) is rapidly transformed into the new hype, following a path similar to Cloud Computing (Armbrust et al., 2010). A general challenge for both researchers and practitioners on answering this issue and meet tight requirements (e.g. time to process), is what kind of design improvements need to be applied and how can the data system in use "*scale*". This requirement for system scalability is applied both in terms of parallel as well as distributed data processing with major architectural changes and use of new software technologies like Hadoop (Apache, 2013a) being the current trend.

DOI: 10.4018/978-1-5225-1759-7.ch021

On the other hand, theoretical definitions of what "Big Data" is and how it can be utilized by organizations and enterprises has been a subject of debates (Jacobs, 2009). On that aspect, the 3V framework has gained considerable attention since it was introduced by Laney (Laney, 2001). In that representation "Big Data" can be defined by three distinctive characteristics, namely: *Volume*, *Variety* and *Velocity*.

The *Volume* represents the ever-growing amount of data, which is generated in today's "Internet of things". On the other hand, the *Variety* of data produced by the multitude of sources like sensors, smart devices and social media in raw, semi-structured, unstructured and rich media formats is further complicating the processing and storage of data. Finally, the *Velocity* aspect describes how fast the data is retrieved, stored and processed. However, dealing with imprecisely defined data formats, growing data sizes, and requirements with varying processing times represent a new challenge to the current systems. From an information processing perspective, the three characteristics together describe accurately what Big Data is. Nonetheless, apart from the 3Vs, which describe the quantitative characteristic of Big Data systems, there are additional qualitative characteristics like *Variability* and *Veracity*. The *Variability* aspect defines the different interpretations that a certain data can have when put in different contexts. It focuses on the semantics of the data, instead of its variety in terms of structure or representation. The *Veracity* aspect defines the data accuracy or how truthful it is. If the data is corrupted, imprecise or uncertain, this has direct impact on the quality of the final results. Both variability and veracity have direct influence on the qualitative value of the processed data. The real value obtained from the data analysis, also called data insights, is another qualitative measure which is not possible to define in precise and deterministic way. A graphical representation of the extended V-Model is given in *Figure 1*.

While the 3V model, shown in Figure 1, provides a simplified framework which is well understood by researchers and practitioners. This representation of the data processes, can lead to major architectural pitfalls on the design of Big Data platforms. A particular issue that should be taken into account, is the

Figure 1. Visualization of the Extended V-Model (adopted from E. G. Caldarola, Sacco, and Terkaj (2014))