Chapter 96
Challenges and Opportunities in Big Data Processing

Jaroslav Pokorny
Charles University, Czech Republic

Bela Stantic
Griffith University, Australia

ABSTRACT

The development and extensive use of highly distributed and scalable systems to process Big Data have been widely considered. New data management architectures, e.g. distributed file systems and NoSQL databases, are used in this context. However, features of Big Data like their complexity and data analytics demands indicate that these concepts solve Big Data problems only partially. A development of so called NewSQL databases is highly relevant and even special category of Big Data Management Systems is considered. In this work we will discuss these trends and evaluate some current approaches to Big Data processing, identify the current challenges, and suggest possible research directions.

INTRODUCTION

Big Data is often characterized by its volume which exceeds the normal range of databases in practice. For example, web clicks, social media, scientific experiments, and datacenter monitoring belong among data sources that generate vast amounts of raw data every day. An interesting characteristic of Big Data is its processing, i.e. Big Data computing. In last years, just Big Data processing is an issue of the highest importance, particularly so called *Big Analytics*. Big Analytics is another buzzword denoting a combination of Big Data and Advanced Analytics. J. L. Leidner (R&D at Thomson Reuters) in the interview with R. V. Zicari (ODMS.org, 2013) emphasizes that buzzwords like "Big Data" do not by themselves solve any problem – they are not magic bullets. He gives an advice how to tackle and solve any problem. There is need to look at the input data, specify the desired output, and think hard about whether and how you can compute the desired result, which is basically nothing but "good old" computer science.

The recent advances in new hardware platforms, methods, algorithms as well as new software systems support Big Data processing and Big Analytics.

DOI: 10.4018/978-1-4666-9840-6.ch096

Effective use of systems incorporating Big Data in many application scenarios requires adequate tools for storage and processing such data at low-level and analytical tools on higher levels. Moreover, applications working with Big Data are both transactional and analytical. However, they require usually different architectures.

Big Analytics is the most important aspect of Big Data computing mainly from a user's point of view. Unfortunately, large datasets are expressed in different formats, e.g., relational, XML, textual, multimedia or RDF, which may cause difficulties in their processing by data mining algorithms. Also, increasing either data volume in a repository or the number of users of this repository requires more feasible solution of scaling in such dynamic environments than it is offered by traditional database architectures.

Clearly, Big Analytics is done also on big amounts of transaction data as extension of methods used usually in technology of data warehouses (DW). Generally DW technology is focused on structured data in comparison to much richer variability of Big Data as it is understood today. Therefore, analytical processing of Big Data Analytics requires not only new database architectures but also new methods for integrating and analyzing heterogeneous data.

Big Data storage and processing are essential for cloud services. This reinforces requirements on the availability and scalability of computational resources offered by cloud services.

Users have a number of options associated with above mentioned issues. For storing and processing large datasets they can use:

- Traditional parallel database systems (shared nothing architectures),
- Distributed file systems and Hadoop technologies,
- Key-value datastores (so called NoSQL databases),
- New database architectures (e.g., NewSQL databases).

In particular, three last categories are not mutually exclusive and can and they should co-exist in many enterprises.

The NoSQL and NewSQL databases present themselves as data processing alternatives that can handle huge volumes of data and provide the required scalability. NoSQL databases are a type of databases which were initiated by Web companies in early 2000s. NewSQL databases are aiming to provide the scale-out advantages of NoSQL databases often on commodity hardware and maintain the transactional data consistency guarantees of traditional relational DBMS. They are also compatible with SQL. Especially, *massively parallel analytic databases* play an important role here. Algorithms supporting Big Analytics are presented on the top of these systems or they are a native part of their implementation.

The chapter is an attempt to cover principles and core features of these systems and to associate them to main application areas of Big Data processing and management in practice, particularly in relation to Big Analytics. We also focus in more extent on challenges and opportunities associated with Big Data.

BACKGROUND

The fundamental concept of generating data has changed recently, in the past, several main sources have been generating data and all others have been consuming data. However, today all of us are both generating data and also consumers of this shared data. Usually we talk about the Big Data when the dataset size is beyond the ability of the current system to collect, process, retrieve and manage the data. (Manyika